Three Articles on a Decision Support System Providing Individual Error Pattern Feedback versus Automated Debiasing of Judgments

Cumulative Dissertation
to obtain the academic degree Dr. rer. pol.
at the Faculty of Business Administration
at the Chair for Business Informatics
of the Catholic University Eichstätt-Ingolstadt

submitted by

Nathalie Balla

Ingolstadt, 2024

Date of the oral exam: 24.10.2023

First advisor: Prof. Dr. Thomas Setzer

Second advisor: Prof. Dr. André Habisch

Contents

1 Overview

The objective of this cumulative dissertation is to mitigate cognitive bias in human expert judgments intending to increase accuracy of judgments. The latter is important as it is known to be essential for business success. The approach for this purpose is to combine the strengths of humans and machines by giving the expert feedback generated by a statistical model (the machine) based on previous errors of the expert. Based on this concept of collaborative intelligence, a Decision Support System (DSS) is developed and tested in several experiments. The cumulative dissertation consists of three articles:

Article 1: Balla, N., Setzer, T., Schulz, F. (2023). Feeding-Back Error Patterns to Stimulate Self-Reflection versus Automated Debiasing of Judgments. *Proceedings of the 56th Hawaii International Conference on System Sciences.*

Article 2: Balla, N. (2023). A Decision Support System Including Feedback to Sensitize for Certainty Interval Size. Forthcoming in *Operations Research Proceedings 2022.*

Article 3: Balla, N., Setzer, T. (2023). Debiasing Judgmental Decisions by Providing Individual Error Pattern Feedback. Submitted to *Decision Support Systems.*

Article 1 and Article 2 investigate different aspects of the DSS and Article 3 merges and supports the statements thereof. Article 1 considers the impact of personal error pattern feedback on further point estimates. The error feedback is based on personal prior judgments originating from different categories, assuming that experts selectively apply the feedback and are able to reduce bias and error. Thereby it is examined, how the feedback is used to change the direction of error and to reduce bias and error. This is investigated in general disregarding categories as well as selectively regarding difference between categories. Article 1 also covers the comparison between human corrected bias and machine auto-corrected bias. Article 2 deals with experiments with the same DSS, but focusing on certainty (confidence) interval estimation and decreasing overprecision (overconfidence) and over- and underestimation biases. Here, the DSS requires users to indicate a 90% certainty interval as an answer to estimation questions. It is investigated how feedback based on own error patterns can help to reduce overprecision by broadening certainty intervals. Moreover, aiming to mitigate over- and underestimation biases, shifts of the intervals are examined. Article 3 supports the statements of Article 1 and 2 by taking into account additional experiments with a larger sample size with the same categories as well as new categories to make the results more robust and generally valid. Article 3 also includes a further analysis regarding the comparison between human corrected bias and machine auto-corrected bias.

2 Motivation and Background of the Dissertation

Although DSSs have been employed before, usually assisting decision-making by collecting, displaying, and visualizing relevant information in aggregated form, judgments resulting thereof are nevertheless frequently biased systematically such as by overconfidence, mean bias, or anchoring (Lawrence & O'Connor, 1993; Lim & O'Connor, 1996; Lawrence, O'Connor, & Edmundson, 2000; Lawrence, Goodwin, O'Connor, & Önkal, 2006; Leitner & Leopold-Wildburger, 2011; Blanc & Setzer, 2016).

Similarly, Blanc and Setzer (2015a) find mean and regression biases in cash flow forecasts of experts despite using DSSs. After detecting statistical patterns, they apply auto-correction to these forecasts and find that accuracy can be improved hereby. Auto-correction represents the employment of a statistical method to the expert forecasts to correct them automatically. However, the authors also detect the problem with auto-correction that it corrects all expert estimates regardless of how sure the experts are in generating their estimates. Thus, auto-correction can also lead to suboptimal automatic corrections that result in higher error than the original expert forecast, termed false-correction (Blanc & Setzer, 2015b). In order to avoid this false-correction issue, the authors propose to give experts the opportunity to accept or to overwrite the prediction made by the statistical model depending on their confidence after their judgment and having been confronted with the prediction of the statistical model including a specification of the bias (Blanc & Setzer, 2015b).

Moreover, previous research has found that the highest performance and accuracy in tasks such as forecasting, estimation, or other decision problems is not achieved by either human or machine itself, but in collaboration with each other, termed collaborative intelligence (Haesevoets, De Cremer, Dierckx, & Van Hiel, 2021). In this context, human experts have a greater ability in recognizing and detecting new effects, unseen developments, and structures due to their domain knowledge and intuition. Machines are more consistent, therefore less error-prone, and better at extracting regular patterns from data (Blattberg & Hoch, 1990; Nagar & Malone, 2011; Arvan, Fahimnia, Reisi, & Siemsen, 2019; Zellner, Abbas, Budescu, & A., 2021). Hence, as important company decisions depend on accurate estimations of certain business figures, it is reasonable to incorporate the abilities of humans and machines for judgmental decision-making.

Nevertheless, an important determinant among others for the success of this suggestion is the type of feedback provided to the experts as it is a prerequisite for them to reflect upon the feedback.

The state of the art literature fundamentally differentiates between two types of feedback, namely outcome feedback (OFB) and cognitive feedback (CFB). OFB generally represents information on the exactitude of the given estimate, which can also simply be the correct answer, whereas CFB constitutes information concerning the process and the cause underlying this exactitude (Jacoby, Mazursky, Troutman, & Kuss, 1984). However, OFB in form of only the true answer has been found to be unhelpful in supporting experts in judgment tasks (Remus, O'Connor, & Griggs, 1996; Balzer, Doherty, & O'Connor, 1989; Lawrence et al., 2006). OFB can be useful in a different form, that is as personalized performance feedback as Benson and Önkal (1992) demonstrate in their experiment. They give subjects in the treatment group performance feedback and subjects in the control group no feedback for the task of forecasting the probability of a team to win a football game. CFB, especially in combination with OFB, has shown to have a positive effect on subjects' performance, which is illustrated by Sengupta and Abdel-Hamid (1993), who conduct an experiment in which subjects make staffing decisions after which all subjects received OFB and a separate group additionally received CFB, for instance information on the perceived cost and size of the project.

A specific bias considered is overconfidence as it is one of the most prevalent cognitive biases and often experts have a misleading sense of control driving them to make decisions in which they are overly optimistic, unable to assess their performance (Ancarani, Di Mauro, & D'Urso, 2016). Specifically, this dissertation also takes overprecision into account, which is one of three kinds of overconfidence differentiated by Moore and Healy (2008), besides overestimation and overplacement. Overprecision represents the notion of being overly sure that the own estimate is more accurate than it is in reality. These aspects can lead to poor decision-making, for which reason mitigation of overprecision should be addressed. The first step is its measurement, for example with the help of interval estimation, which is frequently applied in decision analysis. In interval estimation, subjects are asked to indicate, in most cases a 90% confidence interval, to show how certain they are that the true answer lies within this interval. Frequently, decision makers provide intervals where the true answer lies inside the 90% interval in under 50% of times (Soll & Klayman, 2004). Hence, these decision makers seem to be excessively self-confident and unable to determine their own performance and feedback could be helpful for mitigation. In this context, Ancarani, Di Mauro, and D'Urso (2016) found benchmarks to be valuable to provide to experts in order to ease the assessment of their own performance. According to Ancarani et al. (2016), the achievement of reducing overprecision is possible through feedback including prior decisions, which must be given shortly after the first decisions made.

Further, a known challenge is to design the feedback in a way that experts accept it as they are often overconfident in their judgments despite an indication of inferior performance compared to software (Leitner & Leopold-Wildburger, 2011). To overcome this challenge of potential defensiveness, it is required to support a self-reflective process, that is the interpretation and evaluation of own thoughts, emotions, and actions, required for wise decision making (Grant, Franklin, & Langford, 2002; Sasse-Werhahn, Bachmann, & Habisch, 2020). Goodwin (2000) has shown that self-reflection leads to higher performance and greater accuracy enhancement than no self-reflection. In the course of an experiment he gives forecasters statistical information and then requires them to review judgmental predictions. The outcome shows that asking forecasters to provide a statement for the reason why they changed the prediction leads to better results.

In summary, previous research encourages the investigation of machine learned personal error pattern feedback for self-reflection, bias reduction, and accuracy enhancement.

3 Summary of the Cumulative Dissertation

All three articles contribute to the overall objective to achieve bias reduction and accuracy improvement by providing feedback based on own error patterns with a self-developed DSS. An overview of the interrelation of the articles is depicted in Figure 1. Article 1 lays the foundation by testing the effect of the feedback on over- and underestimation as well as accuracy enhancement and compares this to auto-correction. Article 2 considers the impact of the feedback on overprecision and estimation biases in certainty interval estimation. Article 3 underlines the findings of Article 1 and 2 with a larger sample size and new experiments including novel categories. In the following, all three articles will be individually summarized.

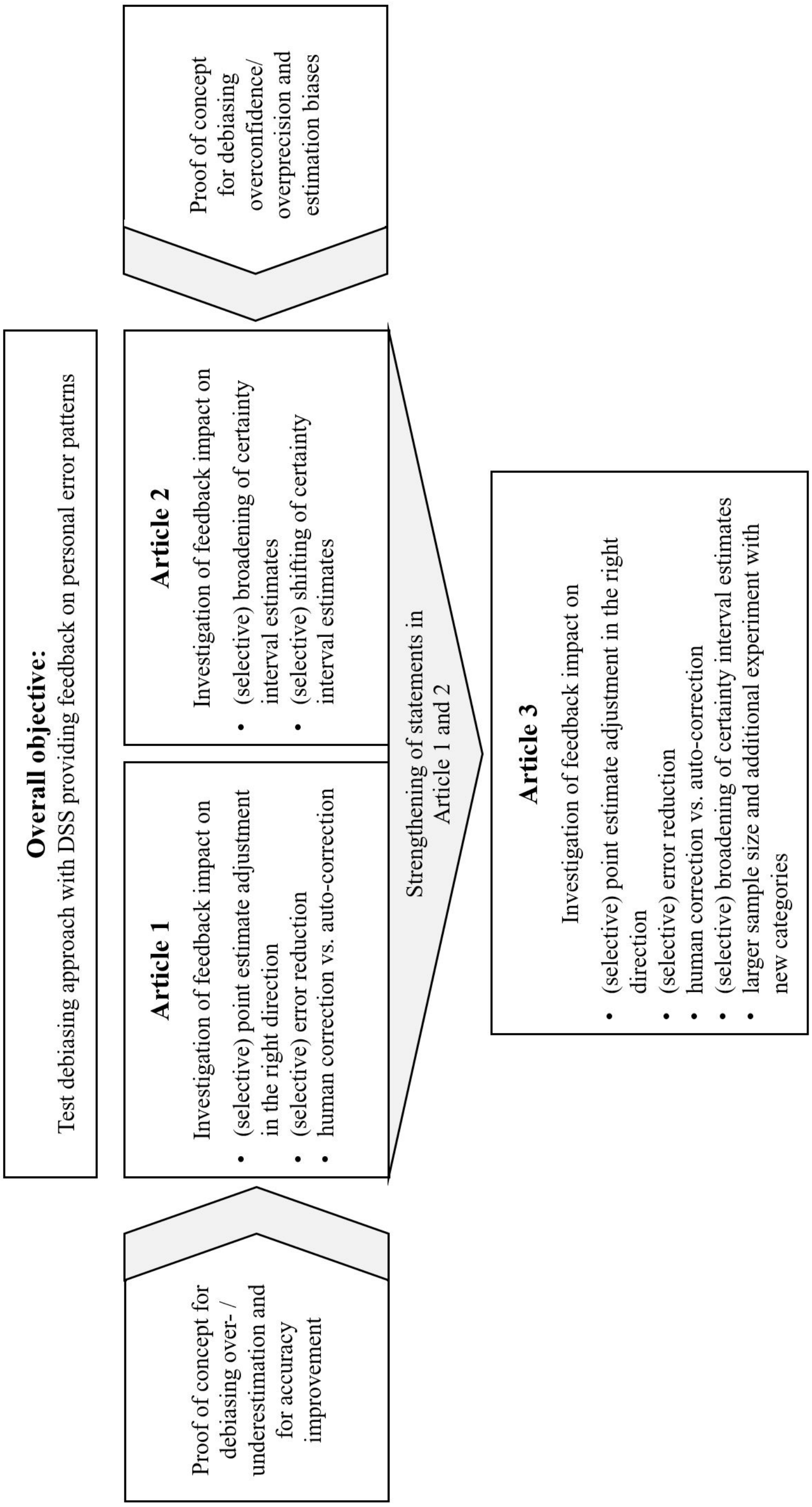

Overall objective:
Test debiasing approach with DSS providing feedback on personal error patterns
Proof of concept for debiasing over-/ underestimation and for accuracy improvement
Article 1
Investigation of feedback impact on
• (selective) point estimate adjustment in the right direction
• (selective) error reduction
• human correction vs. auto-correction
Article 2
Investigation of feedback impact on
• (selective) broadening of certainty interval estimates
• (selective) shifting of certainty interval estimates
Proof of concept for debiasing overconfidence/ overprecision and estimation biases
Strengthening of statements in Article 1 and 2
Article 3
Investigation of feedback impact on
• (selective) point estimate adjustment in the right direction
• (selective) error reduction
• human correction vs. auto-correction
• (selective) broadening of certainty interval estimates
• larger sample size and additional experiment with new categories

Figure 1: Interrelation of the Three Articles

3.1 Article 1: Feeding-Back Error Patterns to Stimulate Self-Reflection versus Automated Debiasing of Judgments

A crucial question in information systems research is how the capabilities of humans and machines can be merged in order to achieve the highest potential performance as it is known that this combination leads to the best results ((Blattberg & Hoch, 1990; Nagar & Malone, 2011; Arvan et al., 2019; Zellner et al., 2021)). Although typical DSSs intend but often fail to reduce human biases, they are still helpful in identifying error patterns. Auto-correction through machine-learned patterns can correct human errors, but in many cases human judgments are corrected erroneously, increasing the error rate needlessly (Blanc & Setzer, 2015b). Targeted feedback of the human's bias may mitigate this issue. As the feedback types personalized performance OFB and CFB have shown to be effective in raising humans' performance in judgments, these feedback forms are also employed in the experiment of this work. Therefore, the approach is to combine the machine's and human's strengths. That is the machine learning the human error pattern and providing it as feedback, and the human detecting new, unseen structures and having implicit knowledge to be able to apply the feedback wisely. This is tested with a DSS in the course of an experiment.

After the configuration of the DSS, developed with Dynamic HTML (PHP) as frontend and a Relational Database Management Server (MySQL) as backend, the configuration items are stored in the database. They involve, for instance, briefing/debriefing, estimation questions including visual cues, and the timing and form of the feedback.

74 subjects participated in this experiment, 34 of them randomized to the treatment group and 40 to the control group. Subsequent to providing subjects with a short briefing, for both groups the experiment starts by asking one question at a time. The general knowledge questions originate from three different categories, namely *number of residents of a country*, *river length*, and *mountain height* worldwide, whereas these are not communicated to subjects for which reason they could equally come up with other, such as regional categories. Example questions are: ”How many residents does France have?”, ”How long is the Hudson River (in km)?”, ”How high is the Mount Everest (in meters)?”. Each question is displayed with a visual cue, intending to provide estimation support and reduce error variance. In case of residents, a map of the respective country including the ten largest cities with an indication of a range of their size is shown. In case of rivers, a map of the river with a scale in the legend and for mountains a topographical map of the mountains with a reference mountain height is depicted.

In addition to answering the questions with a point estimate, subjects are also required to indicate a 90% certainty interval, in which they are 90% certain that the correct answer lies within that range.

The experiment consists of two sequences with each 15 questions with five questions out of each category, where the treatment group receives feedback between these two sequences and the control group receives a blank page inviting for a break, both lasts for 30 seconds. Here, the combination of OFB and CFB refers to the mean percentage error (MPE), reflecting a potential mean bias of the subject, the given and correct answer to each previous question, and an indication if the provided certainty interval by the subject includes the correct answer per question. The individual correct answers per question in the feedback provide additional information, which categories drive the MPE and in which of them over- or underestimation occurs to enhance reflection on the manner of adaptation of estimations. It is noted that feedback is strictly related to patterns in a subject's own error history. The MPE is chosen as it is well comprehensible, although its application to the following sequence of questions is not trivial for subjects. The MPE intends to be a simple example of application as a proof of concept for feedback that is learned by a statistical model, hence many alternative models can be applied.

The MPE would ideally be applied as follows. If a subject receives a MPE of 50%, it means that their estimates surpass the correct answers by 50% on average and correction would mean to take $\frac{2}{3}$ of the next estimates. The intention of the feedback is to confront the subjects with their own error pattern, make them aware of a potential mean bias and invite them to reflect on it. After contemplation, subjects may try to correct their bias, for example by applying the feedback in a category-specific manner. These categories or structures that a human may be able to recognize, are most likely not identifiable by a machine. The idea is that subjects must make new estimations with the help of the generic feedback, which is calculated across all of their past answers, regardless of categories, and thus has to be applied wisely. For example, a mountaineer likely has great knowledge about mountains and would probably estimate mountain heights rather precisely also before the feedback and afterwards would apply the feedback less profound to that category due to their awareness of this category-specific knowledge.

After the feedback and the blank page, both treatment groups are asked to answer another 15 questions, which are new, unseen questions but from the same categories. At the end of the experiment, both treatment groups are provided with the same feedback as well as demographic questions.

Additionally, the mean absolute percentage error (MAPE) is computed as a performance measure between the sequences and in total, also because payouts are based thereon. That is, every subject receives a payout for participation and has the opportunity to win an additional monetary prize, where chances are higher the lower the MAPE is in order to incentivize subjects.

The experiment is meant to imitate experts making judgments in their area of responsibility as they have general knowledge in this field, similar to the subjects who are assumed to have basic knowledge about the general knowledge questions. Moreover, subjects receive visual cues as experts also have additional information available for their estimations. Just like experts have their strengths and weaknesses in specific subfields and may be more biased in some than in others, the expectation is that subjects perform better in certain categories than in others.

In the analysis of the experiment five sub-assumptions (hypotheses) are tested. They are meant to underline the key assumption that a personalized error pattern can foster wise and selective deliberation and application of the feedback and thus reduce bias and increase accuracy. All of these sub-assumptions are supported by stronger results in the treatment group compared to the control group, of which three are significant at a 10% significance level.

They state the following. Subjects receiving MPE-feedback adjust their MPE of the first question sequence in the right direction after the feedback, in general (1) and in a category-specific manner (2), more often than without feedback. An example for a change in the right direction is, if a subject has an MPE of -20% in the first sequence, suggesting underestimation, and an MPE above -20% after the feedback, the subject most likely accepted the feedback to counteract underestimation. Subjects receiving MPE-feedback reduce their MAPE from the first to the second sequence, in general (3) as well as category-specific (4), more often compared to without feedback. The application of auto-correction in the control group compared to subjects applying the feedback in the treatment group leads to less improvements of MAPE in the second sequence (5). The auto-correction is computed by taking the answers of the second sequence of the control group and calculating the auto-corrected answers with the help of the MPE.

By demonstrating that subjects in the treatment group are able to change the MPE in the right direction after the feedback, it is already shown that they indeed reflect on the feedback and integrate it into their further estimations. This is also the case for the category-specific application, which illustrates that the majority of subjects in the treatment group is able to selectively apply the feedback, stronger to categories where it is more necessary. Moreover, the treatment group shows the capability to reduce their MAPE after the feedback, generally as well as selectively, meaning to reduce error the most where it is the highest, which again leads to an overall accuracy increase. In addition, the self-correction of the subjects in the treatment group with the help of the feedback leads to a higher ratio of accuracy improvements compared to auto-correction in the control group, especially in those categories where the MAPE is high in the first sequence. For this reason we can claim that this is an appropriate approach to mitigate the false-correction problem as well as to reduce bias and error in estimation tasks by implementing collaborative intelligence.

3.2 Article 2: A Decision Support System Including Feedback to Sensitize for Certainty Interval Size

One of the most dominant cognitive biases influencing decisions - mostly negatively - is overconfidence. Previous research found that overconfidence occurs frequently in experts' judgments, which may lead to decreasing performance and detrimental judgments (Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Ancarani et al., 2016; Shipman & Mumford, 2011). Precisely, this work considers overprecision denoting an excessive certainty about one's estimate being closer too the correct answer than it actually is (Moore & Healy, 2008). By examining inventory decisions Ren and Croson (2013) detected that overprecision can lead to an underestimation of the variance of demand. The authors measure overprecision by requiring subjects to provide 90% confidence intervals as an answer for every general knowledge question they are asked. This confidence interval indicates a subject's certainty of 90% that the correct answer lies inside this range. In the case that the true answer lies inside the given interval nine out of ten questions, the subject is found to be well calibrated.

A similar approach to measure overprecision is used in the research of Klayman (1999). He requires subjects to give a numerical estimate together with a 90% confidence interval depicted by an upper and lower bound, where the probability is 90% that the true answer lies between the bounds. Regarding the aim of decreasing overprecision, Ancarani et al. (2016) found a promising approach with the help of feedback based on previous decisions and given shortly after judgment.

Therefore, we conduct the same experiment as in Article 1, with 20 subjects in the treatment and 21 subjects in the control group, but with focus on reduction of overprecision in certainty interval estimation. Here, we analyze how the feedback including the MPE and information per question if the correct answer lay inside the 90% certainty interval or not leads to a decline in overprecision and estimation bias.

Two main considerations are investigated. First, the broadening of the average interval between the first and second sequence. This is compared between the treatment group that receives the feedback described above and the control group that receives no feedback. The broadening of the intervals is examined generally and category-specifically to find selective feedback integration. Category-specific means that we examine how often the interval was broadened the most after the feedback in the same category in which the correct answer lay outside the interval most frequently before the feedback. The categories are the same as in Article 1. The assumption is that the broadening may result from the feedback on the correct answer lying within the 90% certainty interval or not.

The second main consideration is the shifting of the intervals between the two sequences compared between treatment and control group, which is also studied generally and category-specifically. If both of the average bounds move in the direction of the MPE feedback, that is if the MPE is negative, the bounds should be larger than before the feedback and vice versa, then it can be assumed that the subject reduced their estimation bias by shifting their interval. If shifts occur, they may originate from the MPE feedback, as this points in a specific direction.

The presumption is, overprecision is present when the bounds are set too narrow and the interval does not include the correct answer, because the subject then is overly certain that their estimation is close to the correct answer, which is not the case. If the 90% certainty interval is broadened after the feedback in case it was too narrow to include the correct answer before the feedback, this is an indication that the respective subject became aware of their overprecision through the feedback, reflected on it, applied the feedback and reduced the overprecision.

Concerning the shifting, the presumption is that subjects hold an estimation bias, that is over- or underestimation, in case both, the upper and lower bound of their provided interval, lie above or below the correct answer. In case the upper and lower bound are shifted upwards or downwards in the direction of the MPE feedback, so that the correct answer lies within the interval, it can be expected that the respective subject internalized the given feedback. Subsequently, the subject integrated the feedback in further estimations to reduce the estimation bias. If the subjects in the treatment group are able to adjust their certainty intervals after the feedback in a differentiated manner, for example to broaden the interval most in that category, where the correct answer lay within the interval the least before the feedback, they indicate the capability to recognize novel structures. They also show to be able to then selectively and wisely apply the feedback accordingly to reduce overprecision and estimation bias.

The results support all of the assumptions as the treatment group continuously shows a stronger performance than the control group. Not all results are significant, most likely due to the small sample size. However, it is also demonstrated that the reduction in overprecision and estimation bias leads to an overall error decrease. Hence, using a DSS providing feedback based on personal error patterns is suitable for debiasing overprecision and estimation bias in interval estimation and a worthwhile approach to pursue in further research.

3.3 Article 3: Debiasing Judgmental Decisions by Providing Individual Error Pattern Feedback

Article 3 merges and supports the aspects of Article 1 and Article 2. That is, presenting a novel DSS and testing the impact feedback based on personal error patterns learned by a statistical model. Further, this article outlines the potential of the DSS as a general functional system for expert judgments.

Underlining the general functionality of the DSS, additional data from a further experiment, an additional analysis for the comparison between human corrected judgments with feedback and auto-corrected judgments, as well as another experiment considering three new categories, are provided.

For the additional experiment with the same categories as above, *number of residents of a country*, *river length*, and *mountain height*, the sample consisted of 97 students, where 51 subjects were in the treatment and 46 in the control group. The results of this experiment support all of the seven hypotheses considered in this work as the treatment group's performance is constantly stronger. Four out of seven hypotheses of this paper are significant at a 5% level.

The hypotheses state the following. Subjects that receive feedback adjust their MPE of the first sequence in the right direction after the feedback more often than the control group (1). Subjects receiving feedback adjust their MPE of the first sequence in the right direction after the feedback the strongest in the category where the MAPE was the highest in the first sequence, more often than the control group (2). The treatment group reduces their MAPE from the first to the second sequence more often than the control group (3). The treatment group reduces their MAPE in the second sequence the most in the category in which the MAPE was the greatest in the first sequence, more often than the control group (4). The application of auto-correction in the control group compared to subjects applying the feedback in the treatment group leads to a smaller number of improvements of MAPE in the second sequence (5). Subjects in the treatment group broaden their average interval after the feedback, in case it did not include the correct answer a certain number of times before the feedback, more often than the control group (6).

This is also considered for category-specific application of feedback to the average interval size (7).

Regarding the additional analysis for the comparison between human corrected judgments with feedback and auto-corrected judgments, it is shown that in case the machine would know the categories and be able to correct in a category-specific way, there is strong improvement compared to the machine not correcting category-specifically. However, the auto-correction still exhibits lower performance than the human using the feedback.

The new categories in the additional experiment are *beeline distances between cities worldwide*, *number of calories in a certain food*, and *heights of famous buildings*. The visual cues for these categories are respectively a map showing the distance between the two cities without a scale but with a hint of the beeline distance between Berlin and Paris, a nutrition table excluding the calories, and a picture of the respective building next to the statue of liberty or a one family house with its height as reference. 32 of 61 subjects participated in the treatment group and 29 in the control group. For this experiment with the new categories, we conducted the same analysis and found strong support for all hypotheses except for (6) and (7), which may be due to the small sample size.

Overall, the additional experiments and analyses underline the majority of results of the experiments in Article 1 and 2. Further, using other categories makes the results independent of categories and therefore more generally valid. Thus, there are many indications that humans are able to self-reflect on personalized error pattern (MPE) feedback including information about errors on the single questions to increase accuracy and reduce over- and underestimation and overprecision. Moreover, humans are able to recognize unseen patterns and then apply the provided feedback selectively and wisely such that the largest errors are decreased most. This leads to a higher error reduction than auto-correction can achieve as a statistical model is not able to recognize the categories and other structures in that human-specific manner.

4 Details on the Articles

Article 1: Feeding-Back Error Patterns to Stimulate Self-Reflection versus Automated Debiasing of Judgments
Authors: Nathalie Balla, Thomas Setzer, Felix Schulz
Publication: Proceedings of the 56th Hawaii International Conference on System Sciences (online)
Conference Presentation: 56th Hawaii International Conference on System Sciences

Contribution of Authors to Specific Tasks			
	Nathalie Balla	Thomas Setzer	Felix Schulz
Literature Review	70%	20%	10%
Experimental Design	60%	30%	10%
Development of Experiment Tool	60%	30%	10%
Experiment Execution	70%	20%	10%
Data Analysis	70%	20%	10%
Writing of Paper	60%	20%	20%

Table 1: Contribution Shares of Authors of Article 1

Article 2: A Decision Support System Including Feedback to Sensitize for Certainty Interval Size
Authors: Nathalie Balla
Publication: Forthcoming in Operations Research Proceedings 2022
Conference Presentation: International Conference on Operations Research - OR 2022

Article 3: Debiasing Judgmental Decisions by Providing Individual Error Pattern Feedback
Authors: Nathalie Balla and Thomas Setzer
Publication: Submitted to *Decision Support Systems*
Conference Presentation: -

Contribution of Authors to Specific Tasks		
	Nathalie Balla	Thomas Setzer
Literature Review	80%	20%
Experimental Design	60%	40%
Development of Experiment Tool	70%	30%
Experiment Execution	90%	10%
Data Analysis	80%	20%
Writing of Paper	70%	30%

Table 2: Contribution Shares of Authors of Article 3

5 Future Research

This dissertation is embedded into a research plan, where the presented experiments relate to the first out of three different scenarios. Scenario two and three are meant to be investigated in experiments in future research.

The first scenario deals with situations of low complexity for the human and high complexity for the machine. In this condition there are latent topics (categories). Humans are easily able to identify them, but machines cannot as they are only able to offer aggregated feedback and auto-correct judgments equally. The results of the dissertation show that humans can reflect on their personal error pattern and know in which categories they may be biased leading to a higher performance as they know if and how to incorporate the feedback.

The second scenario considers situations with low complexity for human and machine. Thus, the intention is that the latent topics are known by both, human and machine and auto-correction can be done selectively on topics when the machine is provided with more data thereon to learn them. The input of the auto-correction by the machine may profit from the information about latent topics as human biases may be specific to topics.

The third scenario takes into account situations with high complexity for human and machine. Latent topics may be overlapping and questions not distinctly assignable. Similar to the first scenario, the machine will not be able to identify structures in these questions and give specific feedback. However, the assumption is that the human can nonetheless apply the general feedback wisely and potentially selectively, because he or she possesses domain knowledge and intuition to detect latent topics.

Overall, this dissertation is able to illustrate great potential for the DSS providing personalized error pattern feedback to human experts to mitigate biases and increase accuracy of estimations. It contributes to the information systems literature in that it offers insight on how humans use feedback that stems from their own errors and how collaborative intelligence can be achieved.

6 Abstracts

6.1 Feeding-Back Error Patterns to Stimulate Self-Reflection versus Automated Debiasing of Judgments

Abstract

Automated debiasing, referring to automatic statistical correction of human estimations, can improve accuracy, whereby benefits are limited by cases where experts derive accurate judgments but are then falsely "corrected". We present ongoing work on a feedback-based decision support system that learns a statistical model for correcting identified error patterns observed on judgments of an expert. The model is then mirrored to the expert as feedback to stimulate self-reflection and selective adjustment of further judgments instead of using it for auto-debiasing. Our assumption is that experts are capable to incorporate the feedback wisely when making another judgment to reduce overall error levels and mitigate this false-correction problem. To test the assumption, we present the design and results of a pilot-experiment conducted. Results indicate that subjects indeed use the feedback wisely and selectively to improve their judgments and overall accuracy.

Keywords: decision support system, debiasing, automated debiasing, feedback, self-reflection

6.2 A Decision Support System Including Feedback to Sensitize for Certainty Interval Size

Abstract

In decision-making overconfidence and estimation biases can lead to sub-optimal outcomes and accuracy loss. A debiasing strategy presented in this work is to use feedback based on the error pattern of own previous absolute and 90% certainty (confidence) interval estimates. This is comprised in a decision support system (DSS) and applied in an experiment, where results indicate support for the key assumption that subjects are able to selectively reduce their overconfidence and their estimation bias, if present, with the help of the provided feedback.

Keywords: DSS, interval estimation, overconfidence, overprecision, debiasing

6.3 Debiasing Judgmental Decisions by Providing Individual Error Pattern Feedback

Abstract

We present a novel Decision Support System (DSS) that provides experts with feedback on their personal potential bias based on their previous error pattern. As the feedback stems from an expert's own error pattern, it intends to enhance their self-reflection, foster wise consideration of the feedback and mitigate potential biases. Common DSSs, which typically filter and visualize relevant information, also aim to support rational decisions. However, experts using such DSSs generally still exhibit systematically flawed decisions. These systematic errors can be detected and automatically corrected by DSSs, respectively machines, by correcting humans' judgments with a statistical method afterwards. Nevertheless, this often leads to suboptimal outcomes, because the machine also falsely corrects originally proficient judgments as a machine is, unlike a human expert, not conscious of implicit knowledge or possibly unaware of structural breaks. We assume that experts are able to apply the above described feedback systematically and selectively to different decision tasks and to therefore reduce their potential bias and error. To test this assumption, we conduct experiments with the DSS. Therein, subjects provide point estimations as well as certainty intervals and subsequently receive feedback, which is given by a machine that learns the error pattern of the respective subject based on previous answers. After the feedback, subjects answer further questions. Results indicate that subjects reflect on their own error pattern and apply the feedback selectively to further estimations to reduce overall bias and error.

Keywords: Decision Support System, Debiasing, Feedback, Self-Reflection, Auto-Debiasing, Collaborative Intelligence

References

Ancarani, A., Di Mauro, C., & D'Urso, D. (2016). Measuring overconfidence in inventory management decisions. *Journal of Purchasing and Supply Management*(22(3)), 171-180.

Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega*(86), 237-252.

Balzer, W. K., Doherty, M. E., & O'Connor, R. J. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*(106(3)), 410-433.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*(8), 559-573.

Blanc, S., & Setzer, T. (2015a). Analytical debiasing of corporate cash flow forecasts. *European Journal of Operational Research*(243(3)), 1004-1015.

Blanc, S., & Setzer, T. (2015b). Improving forecast accuracy by guided manual overwrite in forecast debiasing. In *Twenty-third european conference on information systems (ecis)* (p. Paper 66).

Blanc, S., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*(69), 3951-3962.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*(36(8)), 887-899.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*(16), 85-99.

Grant, A., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality*(30(8)), 821-836.

Haesevoets, T., De Cremer, D., Dierckx, K., & Van Hiel, A. (2021). Human-machine collaboration in managerial decision making. *Computers in Human Behavior*(119).

Jacoby, J., Mazursky, D., Troutman, T., & Kuss, A. (1984). When feedback is ignored: Disutility of outcome feedback. *Journal of Applied Psychology*(69(3)), 531-545.

Klayman, J., Soll, J., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*(79(3)), 216-247.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*(22), 493-518.

Lawrence, M., & O'Connor, M. (1993). Scale, variability, and the calibration of judgmental prediction intervals. *Organizational Behavior and Human Decision Processes*(56), 441-458.

Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*(122), 151-160.

Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – a review of the literature. *European Journal of Operational Research*(213(3)), 459-469.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems*(16(4)), 339-357.

Moore, D., & Healy, P. (2008). The trouble with overconfidence. *Psychological Review*(115(2)), 502–517.

Nagar, Y., & Malone, T. (2011). Making business predictions by combining human and machine intelligence in prediction markets. In *Icis 2011 proceedings* (p. Paper 20).

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*(66(1)), 22-30.

Ren, Y., & Croson, R. (2013). Overconfidence in newsvendor orders: An experimental study. *Management Science*(59(11)), 2502-2517.

Sasse-Werhahn, L. F., Bachmann, C., & Habisch, A. (2020). Managing tensions in corporate sustainability through a practical wisdom lens. *Journal of Business Ethics*(163), 53-66.

Sengupta, K., & Abdel-Hamid, T. K. (1993). Alternative conceptions of feedback in dynamic decision environments: An experimental investigation. *Management Science*(39(4)), 411-428.

Shipman, A., & Mumford, M. (2011). When confidence is detrimental: Influence of overconfidence on leadership effectiveness. *The Leadership Quarterly*(22(4)), 649-665.

Soll, J., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*(30(2)), 299–314.

Zellner, M., Abbas, A. E., Budescu, D. V., & A., G. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*.

A Original Articles

Feeding-Back Error Patterns to Stimulate Self-Reflection versus Automated Debiasing of Judgments

Nathalie Balla
KU Eichstätt-Ingolstadt
nballa@ku.de

Thomas Setzer
KU Eichstätt-Ingolstadt
thomas.setzer@ku.de

Felix Schulz
KU Eichstätt-Ingolstadt
felix.schulz@ku.de

Abstract

Automated debiasing, referring to automatic statistical correction of human estimations, can improve accuracy, whereby benefits are limited by cases where experts derive accurate judgments but are then falsely "corrected". We present ongoing work on a feedback-based decision support system that learns a statistical model for correcting identified error patterns observed on judgments of an expert. The model is then mirrored to the expert as feedback to stimulate self-reflection and selective adjustment of further judgments instead of using it for auto-debiasing. Our assumption is that experts are capable to incorporate the feedback wisely when making another judgment to reduce overall error levels and mitigate this false-correction problem. To test the assumption, we present the design and results of a pilot-experiment conducted. Results indicate that subjects indeed use the feedback wisely and selectively to improve their judgments and overall accuracy.

Keywords: decision support system, debiasing, automated debiasing, feedback, self-reflection

1. Introduction

A key question in current information systems research is how to achieve collaborative intelligence, i.e., how to combine the complementary strengths of machines, which are stronger in extracting regular patterns from data, and humans, which are more adept at considering novel or transferable situations, effects or unseen developments based on domain knowledge and intuition (Blattberg and Hoch, 1990; Nagar and Malone, 2011; Zellner et al., 2021).

We introduce a novel decision support system (DSS) aimed at improving estimation accuracy by fostering collaborative intelligence. The mechanism implemented by the DSS is to feed-back machine-learned personalized error patterns (biases) of an expert to that same expert who then decides how to incorporate that feedback into her or his further judgments.

Accuracy of estimations is vital for enterprises since planning and decision making usually depend on accurate estimations of (future) business figures. As of today, many respective tasks are dominated by judgmental approaches, i.e., by humans with individual backgrounds, attitudes, and estimation heuristics (Klassen and Flores, 2001; McCarthy et al., 2006; Sanders and Manrodt, 2003). A typical DSS supports such tasks by gathering, filtering, and presenting relevant information to derive informed and unbiased judgments.

However, providing additional information does not have an unambiguously positive effect on accuracy and while a huge body of work on DSSs has been published on how to integrate, aggregate, and visualize data to derive accurate estimations and beneficial decision alternatives, empirical evidence shows that the judgments derived by seemingly well-configured DSSs still come out flawed, including biases like overconfidence, mean or regression bias, optimism, over-steering or anchoring (see, for instance, the findings in Blanc and Setzer, 2016; Lawrence et al., 2006; Lawrence and O'Connor, 1993; Lawrence et al., 2000; Leitner and Leopold-Wildburger, 2011; Lim and O'Connor, 1996).

As a recent example derived from a large corporate dataset, Blanc and Setzer (2015a) analyze a set of empirical cash flow forecasts of a multinational corporation, generated by more than one hundred experts from different subsidiaries using forecast DSSs.

URI: https://hdl.handle.net/10125/102800
978-0-9981331-6-4

The authors find that, nevertheless, mean as well as regression biases exist for all business divisions of the company. Furthermore, they find that the statistically identified error patterns allow for an automated statistical correction of the patterns that increases overall accuracy. The authors also show that the estimated model parameters relate to characteristics of the business environments and argue that these provide valuable insights to better understand, quantify, and feed-back presumed biases to the experts to help them to improve the accuracy of future forecasts.

The same authors also show that, since automated correction is applied to estimates regardless of presumably different confidence in the original estimate, appropriate expert expectations are also corrected in the wrong direction. This leads to higher errors than necessary (Blanc and Setzer, 2015b).

To address this problem, for future research the authors suggest a feedback-based DSS that shows the expert, after she or he submitted a forecast, the forecast of a statistical (correction) model together with a description of the bias that might have driven the discrepancy to the expert's expectation. The authors propose to derive such a benchmark forecast by correcting time persistent biases in past expert forecasts. The expert might then be prompted to accept or overwrite the model forecast, ideally overwriting primarily the model predictions that would lead to heavy false-corrections.

The intuition of providing error pattern based feedback and the key assumption of such an approach that experts are capable to consider the error-feedback wisely and selectively seems compelling. However, this key assumption has, to our knowledge, not been tested so far. For instance, when an estimation task falls in a domain the expert is very familiar with and is sure that the error-feedback is likely not to apply to his or her current judgment, it should be neglected. In cases where an expert is less confident that no structural bias is at play, the feedback might be accepted and the estimation adjusted. Overall, an expert must be capable to make informed decisions if the structural error pattern he or she received is likely to be valid (i.e., whether a bias might indeed be at play).

We present the architecture of a novel DSS together with the design and the results of a first experiment to test this assumption. The DSS addresses the problem that auto-debiasing of experts' judgments leads to decreasing accuracy if the expert made the judgment knowledgeably and accurately, but the model falsely corrects it. The DSS design further aims at providing guidance on how to systematically improve further judgments, i.e., to learn based on errors made in the past. Such a type of DSS may be important for several fields in business, where decision-makers are dependent on the accuracy of estimations and predictions.

The experiment is the first in a series of experiments currently conducted to find evidence for such wise and systematic adjustments after receiving personalized error patterns as feedback, and whether this leads to error reduction. In the experiment, subjects are asked to estimate quantities from different general knowledge categories, while categories are not communicated, and error-feedback in terms of their mean bias (measured as mean percentage error, MPE) is displayed after a sequence of estimations made.

The experiment is designed to make the key assumption described above testable by few sub-assumptions (hypotheses) related to changes of the MPE in the right direction after feedback, whether change is emphasized in categories with higher before-feedback MPE, and whether accuracy improvement is achieved compared to subjects not receiving the feedback, with and without auto-correction of their estimates. Results indicate that subjects indeed seem to use the feedback wisely and selectively to improve judgments.

The rest of this article is organized as follows. In Section 2, we review previous research on auto-debiasing and feedback-based DSS with regard to whether they hint at specific feedback mechanisms promising to enable wise and selective consideration of error-correction feedback. In Section 3, we describe the DSS used as the experimental infrastructure. In Section 4, we present the design and the results of a first experiment that serves as a general proof of concept for the DSS. In Section 6, we discuss the results of our work so far, conclude, and outline future research on error feedback-based DSS.

2. Prior Work on Bias-Related Feedback vs. Auto-Correction

We start reviewing findings with auto-correction, and then review approaches to foster debiasing using feedback. Finally, we discuss their suitability to foster learning, improve judgment accuracy and mitigate the false-correction problem inherent with auto-correction.

As aforementioned, Blanc and Setzer (2015b) discuss accuracy gains through auto-debiasing, referring to the automatic correction of experts' forecasts by a statistical model learned on previous experts' errors. Figure 1 shows the distributions of absolute percentage error (APE) improvements of forecasts when using the corrected forecasts instead of the original expert forecasts per decile of the confidence interval around the correction model's forecast. The larger the correction, the higher the variance of error differences. The most deviating decile bins contain the heaviest accuracy gains and losses.

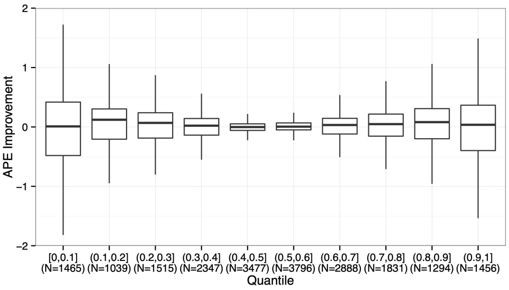


Figure 1. APE distribution by decile in the confidence interval of the auto-corrected forecast Blanc and Setzer, 2015b

The authors argue that, as auto-correction is applied to estimates regardless of confidence in the original estimate, originally sound expectations are also corrected in the wrong direction, leading to high errors specifically in outer decile bins. Hence, they suggest to prompt experts to accept or overwrite a model forecast if the expert forecast exceeds certain confidence bounds, as in cases of extreme deviations either a strong bias might be at play or the expert forecast might be based on specific knowledge and indeed be appropriate.

This perspective has merit as experts can be assumed to have higher confidence and knowledge in certain estimation tasks and biases are likely to depend on the type of estimation task. However, whether experts are capable of making wise feedback accept/neglect decisions depends on several factors, where one of particular importance is surely the type of feedback provided. Therefore, we now review feedback-based DSSs and whether they appear promising for the task of making wise error-feedback consideration decisions.

A common distinction of feedback types is outcome feedback (OFB) and cognitive feedback (CFB). OFB refers to "information that describes the accuracy or correctness of the response" (Jacoby et al., 1984, p. 531), and is often solely the correct answer. CFB is "information regarding the how and why that underlies this accuracy" (Jacoby et al., 1984, p. 531).

Regarding outcome feedback, Remus, O'Connor, and Griggs (1996), Balzer, Doherty, and O'Connor (1989) and Lawrence, Goodwin, O'Connor, and Önkal (2006), amongst others, show that OFB in form of providing correct answers is rather ineffective, and many studies question the usefulness of OFB of that type in general (Balzer et al., 1989). It is argued that such information is insufficient to improve judgment. It has even been shown that better performing experts avoid using OFB of that type (Lawrence et al., 2006; Remus et al., 1996).

In contrast, OFB in the form of personalized performance feedback seems more suitable. As an example, Benson and Önkal (1992) studied performance feedback in probability estimation. In their experiment, subjects made four weekly predictions of football games for the following weekend regarding the probability for a team to win. Subjects of the treatment group received performance feedback while control group subjects did not. The authors find that performance feedback helped to increase forecasting accuracy.

Fischer and Harvey (1999) observed that feedback originating from performance on one trial increases motivation of the subject in the next trial. Here, subjects were asked to combine sales forecasts of others, where the treatment group received feedback on their first trial before the their second trial. The feedback showed the own forecast, the actual outcome, and the respective error. The results indicate that such feedback does help to learn and also induces motivation through goal-setting as the feedback functions as a goal to outperform.

Although we do not find studies focusing on selective incorporation of feedback and adjustment of estimations, based on prior research, actionable error-feedback seems to be a promising candidate for our setting.

Concerning cognitive feedback, Sengupta and Abdel-Hamid (1993) published an article in Management Science that presents an experiment integrating CFB in DSSs. 47 subjects performed tasks as project managers in terms of staffing decisions for a software project, which involved trade-offs between cost and time plan. After making decisions, every subject received outcome feedback in terms of a report on the current stage of the project. In the CFB group, CFB was available in form of task information through plots of variables over the project's life span (such as information on the perceived cost and size of the project) and a summary of the past interval. Experimental results show that subjects with access to CFB (in addition to OFB) performed best compared to the group receiving only OFB.

Sengupta (1995) conducted further experiments, where subjects had to conduct personnel screening. The treatment group received OFB as well as CFB whereas the control group received OFB only. OFB was shown as the rating decisions made by the expert commitee and CFB as the committee's decision strategy regarding similar jobs as well as consistency scores and information on a subject's own decision strategy. The findings show that subjects receiving OFB together with CFB tend to outperform those receiving OFB only. Combining performance with cognitive feedback therefore seems like an approach worthwhile to be pursuit for our purpose.

However, a severe challenge is the acceptance of feedback by an expert in general, as it has been found that experts are usually overconfident in their own expectations even if their ability is shown to be inferior to the estimate provided by software (Leitner and Leopold-Wildburger, 2011). Therefore, a challenge is fostering a self-reflective process, i.e., the interpretation and assessment of own thoughts, emotions, and actions, required for directed change and key to wise decision making (Grant et al., 2002; Sasse-Werhahn et al., 2020).

For instance, in an experiment by Goodwin (2000), prompting forecasters to revise judgmental forecasts after statistical information has been provided did not improve accuracy, whereas asking forecasters to adjust a forecast while requiring reflection by providing a reason for the adjustment performed best. It has also been found that specifically feedback like error-feedback drives reflective processes, which in turn affects if and how the feedback is accepted and used. For example, Sargeant, Mann, van der Vleuten, and Metsemakers (2009) conducted interviews with physicians who evaluated assessment feedback they received. This reflection was useful in terms of how to apply the feedback.

Overall, in search for a promising feedback-type, previous work on feedback, debiasing, and self-reflection encourages the usage of an expert's own error pattern – relating to a potential bias that can be understood and corrected – as performance related feedback type. In addition, it seems suitable to induce self-reflection as it is different to external feedback often adopted insufficiently. Thereby we provide both, promising types of OFB and CFB.

3. Experimental Infrastructure and Procedure

We now introduce the DSS infrastructure used for our experiments from a procedural perspective together with key considerations, while keeping technical details short. We illustrate several components by providing examples of their implementation in the first experiment.

Technically, the DSS is developed as a Web-App using Dynamic HTML (PHP) as frontend, and a Relational Database Management Server (MySQL) as backend containing the parameterization of the experiments, storing outcomes, and used for analyzing the answers and reactions of the subjects. The error pattern derivation, its presentation as feedback as well as the calculation of loss functions are provided by tools written in PHP and *R*.

An overview of the steps supported by the DSS is depicted in Figure 2. First, an experiment is configured using a Web-based tool and the configuration is stored in a database. Configuration items are pages for briefing/debriefing, comprehensibility questions, estimation questions to be answered by treatment and control group subjects, the loss function that measures performance, rules when feedback of what form is provided, texts and visuals provided with a question, rules when an experiment terminates, and a final questionnaire form.

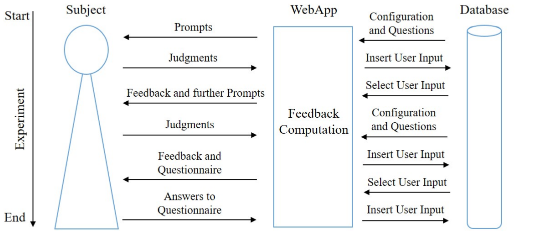
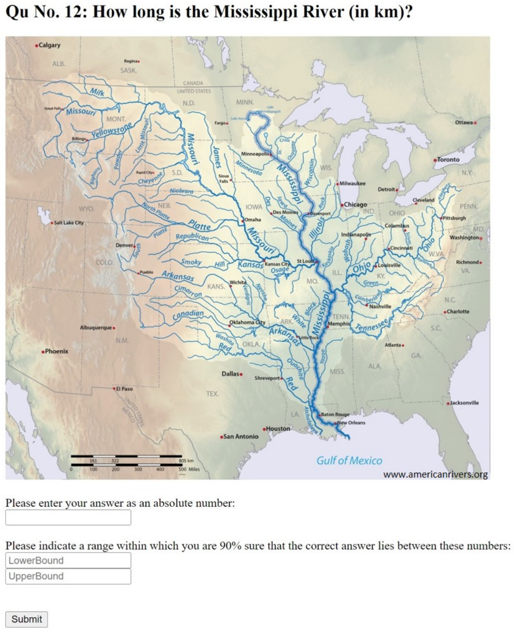


Figure 2. Experimental Infrastructure and Processes

Then, the subjects are randomly assigned to the treatment or control group, shown information on the experiment, asked comprehensibility questions and the experiment itself starts by prompting for judgments. An example prompt is shown in Figure 3. Here, the task is



Qu No. 12: How long is the Mississippi River (in km)?

Please enter your answer as an absolute number:

Please indicate a range within which you are 90% sure that the correct answer lies between these numbers:

LowerBound

UpperBound

Submit

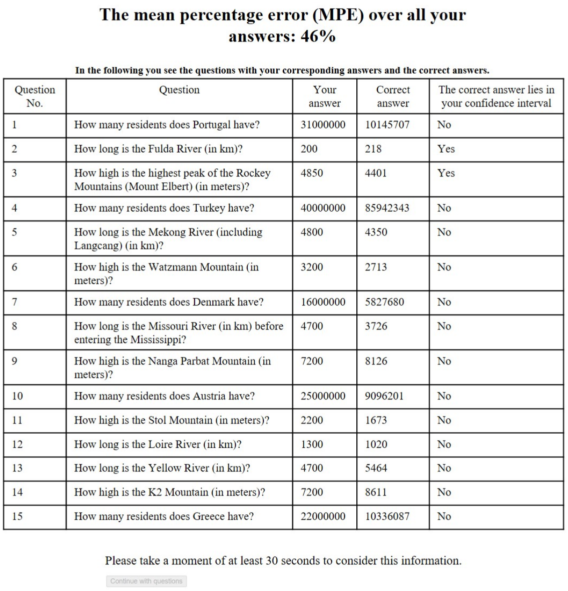
Figure 3. DSS Interface – User Prompt Example

to estimate the length of the Mississippi in kilometers, where guidance is provided by a map and a legend indicating the scale. A subject is also asked to indicate her or his 90% confidence interval – the interval to which a subject is 90% sure that the correct answer lies within. After the answer is submitted, the next prompt is displayed, and after a defined number of questions either feedback is provided for 30 seconds (treatment group) or a blank page is shown inviting to take a 30 second break (control group). In case feedback is given, the subject's error pattern (bias) is computed and displayed together with her or his individual estimation errors.

In our first experiment, the subject's mean bias computed as her or his mean percentage error (MPE) across all the previously given answers, is shown as feedback. As this DSS is meant to demonstrate general functionality as a proof of concept, the MPE is only a simple example of a statistical model. Other models can be used to calculate indications of other biases. MPE is computed as follows: per estimate made, the difference between the estimate and the actual (the actually correct answer) is computed and that difference is divided by the actual and multiplied by 100. MPE is then the mean of these values and therefore also calculated across all categories. MPE is chosen for reasons of comprehensibility and ease of applicability for debiasing. For example, a MPE of 0.5 means that estimates exceed actuals by 50% on average, and correction means to take only $\frac{2}{3}$ of a further estimate. Figure 4 shows an example feedback page with the (potential) mean bias of the subject.

The intention of the feedback is to make a subject aware of a potential mean bias, possibly derived by previously given answers. A potential cognitive bias might be mentally corrected by a subject when providing further novel questions. This may be category-specific, although categories are not mentioned or used by the DSS. Thus, the aim of the feedback is to make subjects reflect on previous error patterns to improve future estimations. Hence, the subjects must make novel estimations applying the generic feedback that is computed across all their previous answers and needs to be cognitively wisely applied.

After the feedback or the blank page, a subject is faced with another sequence of novel judgments from the same categories and the experiment terminates with a final feedback and a user survey.


The mean percentage error (MPE) over all your answers: 46%

In the following you see the questions with your corresponding answers and the correct answers.

Question No.	Question	Your answer	Correct answer	The correct answer lies in your confidence interval
1	How many residents does Portugal have?	31000000	10145707	No
2	How long is the Fulda River (in km)?	200	218	Yes
3	How high is the highest peak of the Rockey Mountains (Mount Elbert) (in meters)?	4850	4401	Yes
4	How many residents does Turkey have?	40000000	85942343	No
5	How long is the Mekong River (including Langcang) (in km)?	4800	4350	No
6	How high is the Watzmann Mountain (in meters)?	3200	2713	No
7	How many residents does Denmark have?	16000000	5827680	No
8	How long is the Missouri River (in km) before entering the Mississippi?	4700	3726	No
9	How high is the Nanga Parbat Mountain (in meters)?	7200	8126	No
10	How many residents does Austria have?	25000000	9096201	No
11	How high is the Stol Mountain (in meters)?	2200	1673	No
12	How long is the Loire River (in km)?	1300	1020	No
13	How long is the Yellow River (in km)?	4700	5464	No
14	How high is the K2 Mountain (in meters)?	7200	8611	No
15	How many residents does Greece have?	22000000	10336087	No

Please take a moment of at least 30 seconds to consider this information.

Figure 4. DSS Interface – Feedback Page Example

MAPE, the mean absolute percentage error, is used as the performance and accuracy criterion to determine improvement or deterioration between question sequences and for the payouts that depend on MAPE values. MAPE is calculated similarly to MPE, but taking the absolute differences between estimates and actuals. Information on how to interpret and apply MPE for debiasing is given in the briefing phase (without telling subjects that they will receive feedback) together with information on MAPE used as performance measure for payouts.

The infrastructure and the scenario characterized above is the one used in our first experiment. The experiment itself will be described in the subsequent section. A broader picture of our research, including other scenarios that will be considered in our research and how the first scenario is embedded in our research plan will be provided in Section 6.

4. Experiment

We first describe the research design in terms of the experiment's configuration (the general experimental procedure including the feedback provided and the loss function is described in Section 3). Second, we present the assumptions explored in the experiment and the measures used to analyze whether we find support for the assumptions. Third, we provide the results.

4.1. Research Design

In the experiment we have 74 subjects (34 in the treatment, 40 in the control group), of which 39 are female and the rest male. 41 subjects are business students and 33 subjects are (school) students.

Subjects are prompted for point estimates of quantities together with a 90% confidence interval from general knowledge categories, namely *number of residents of a country*, *river length*, and *mountain height*. Example questions are: ”How many residents does France have?”, ”How long is the Hudson River (in km)?”, ”How high is the Mount Everest (in meters)?”. The experiment contains two sequences of 15 questions, 30 questions in total. Categories are neither communicated nor used by the correction model, but easy to anticipate by humans.

Estimation tasks are supported by cues: maps of the respective country including the ten largest cities with an indication of a range of their size; maps of the rivers with a scale in the legend; topographical maps of the mountains with a reference mountain height. These visual aids shall reduce error variance, but are also useful for heuristics applied and might trigger specific biases to be recognized as mean bias error patterns. For example, a subject may underestimate the additional river length stemming from the river loops and bends. The subject might then apply the error-feedback to debias her or his estimates only for questions of that type in case other categories seem unbiased.

The scenario mimics experts’ environments where experts have expertise and basic confidence in all categories they are prompted for estimates, consider different types of visual cues and information for different types of estimation tasks, while expertise and heuristics applied might vary amongst categories. A human expert will typically be faced with categories or types of questions where he or she is particularly prone to biases. These types can be human-specific and a machine or statistical method would likely not be able to recognize the same types a particular human might have in mind. For instance, in our experiment it may be that a subject knows river lengths, mountain heights, and population sizes of north and middle European countries well but might be less familiar with other regions and then apply a geographical categorization.

After the first sequence of questions, a subject in the treatment (control) group receives feedback in terms of her or his mean bias measured as MPE (a blank page with the prompt to pause for 30 seconds). The MPE is displayed as inverse performance feedback to the subject, which can be easily applied for debiasing, together with the individual answers given by the respective subject and the actual correct answers per estimation question. The errors per question provide further hints by which categories the MPE might be driven, or where over- or underestimation is identifiable to foster reflection on how to further adapt judgments. We note that feedback is strictly related to patterns in a subject’s own error history.

Following the feedback or the blank page, a subject answers the second sequence of questions, which are completely new to the subject. These questions are from the same categories as used in the first sequence.

After the experiment, a subject receives a debriefing and her or his MAPE is computed. A subject receives a payout for participation and has the chance to additionally win one of two prizes per treatment group. The lower the MAPE of a subject, the higher the chance to receive a prize. This incentivization is meant to increase the motivation and performance of subjects.

The overall experimental procedure is depicted in Table 1. In the following, we will describe the assumptions tested in the experiment.

Table 1. Experimental Design and Procedure

	Treatment Group	Control Group
15 Questions	x	x
Feedback	Yes	No
15 Questions	x	x
Feedback and Demographic Questions	x	x

4.2. Assumptions Studied

We split up the key assumption that one’s own error patterns can foster wise and selective consideration of the feedback, into the (sub) assumptions A1–A5.

A1: MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction.

The direction of change of a subject’s MPE is analyzed to study if a reaction to the feedback can be assumed that leads in the right direction. If a subject receives a negative MPE, her/his subsequent MPE should be less negative or slightly positive and vice versa.

To test A1, per subject we determine the MPE over the answers in the first sequence (before the feedback or blank page) and the answers in the second sequence (after the feedback or blank page). Per subject we then determine whether her or his MPE changed in the right direction, and compare the ratio of right-direction MPE changes in the treatment versus the control group. The assumption is that the ratio is higher in the treatment than the control group and around 50 % in the control group (where no feedback is provided that might cause systematic MPE change).

For A1 we conduct a Fisher’s exact test of independence between the results of the treatment and the control group for the ratio of right-direction MPE changes to detect a significant difference between the proportions of the two categorical variables. The test is one-sided to test if the proportion of cases where the MPE changed in the right direction is higher in the treatment than in the control group. The treatment and control group are independent, relatively small samples, for which reason Fisher's exact test is a suitable non-parametric test.

A2: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MPE, resulting in larger MPE change in this category.

The rational of A2 is that subjects know in which categories they are biased the most and use this knowledge wisely instead of blindly applying the feedback across all categories, as auto-debiasing unaware of categories would do.

To test A2, per subject and category MPE before and after the feedback (or blank page) is calculated. Then, the percentage of matches of the category with the highest absolute MPE in the first sequence and the category with the largest MPE change in the right direction from the first to the second sequence for the subjects in the treatment versus the control group are computed. If the percentage value in the treatment group exceeds the one in the control group, and in the treatment group both categories match in more than $\frac{1}{3}$ of cases (the baseline ratio in case of randomness), selective application of the feedback to specific categories can be assumed.

As for A1, we also test the significance of the difference in the results between treatment and control group with the Fischer's exact test for A2.

A3: MPE-feedback leads to higher MAPE reduction compared to no feedback given.

This assumption differs from A1 as it is related to accuracy improvements as a result of adapted judgment compared to A1 that studies solely MPE changes in the right direction. We note that MAPE might increase although MPE changes in the right direction when changes lead to absolute percentage errors exceeding the MAPE in the first sequence (if the absolute percentage errors increase).

After determining the difference of a subject's MAPE in the first and the second sequence, we calculate the ratio of MAPE improvements of subjects in the treatment versus the control group. We assume this ratio to be higher for the treatment group and again expect a ratio of around 50% in the control group due to random MAPE increases or decreases.

As for previous assumptions, we conduct a Fisher's exact test, again to find a significant difference between the results of treatment and control group.

A4: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category.

The rational of A4 is that subjects selectively apply the feedback to certain categories with high error levels (high bias) such that, respectively, the MAPE declines most strongly in these categories with high MAPE before the feedback.

The MAPE reduction after the feedback or blank page is computed per subject and category to determine the correspondence between the category with the highest MAPE in the first sequence and the category with the largest MAPE reduction from the first to the second sequence. If the ratio in the treatment group exceeds the one in the control group, and in the treatment group both categories match in more than $\frac{1}{3}$ of cases (baseline in case of randomness), wise, category-specific application of the feedback that leads to MAPE reduction can be assumed.

Again, we examine the significance of the difference in the results between treatment and control group by performing a Fisher's exact test.

A5: MPE-error-feedback leads to higher MAPE reduction particularly in categories with high MAPE in the first sequence compared to auto-correction.

Support for this assumption would indicate that feedback-based adjustment can mitigate strong false-corrections inherent when using auto-correction.

To test A5, the percentage of MAPE improvements in the treatment group between sequence one and two is compared to the percentage of hypothetical MAPE improvements through auto-correction in the control group. The improvements by auto-correction in the control group are computed by taking the answers of a subject of the second question sequence and including the MPE of the answers of the first questions sequence in the calculation of all hypothetically corrected answers for the second sequence. Then the MPE over these auto-corrected answers is computed. The Fisher's exact test is again used to test the significance of the difference between the results. Furthermore, per subject we determine the category with the highest MAPE in the first sequence and study whether feedback is beneficial in reducing these high error levels (MAPE) compared to auto-correction. For this specific category we compute how high the improvement is by subtracting the MAPE in the second sequence from the MAPE in the first sequence and taking the average per group thereof. We assume this value to be higher in the treatment group versus the control group (with auto-correction).

To test the significance of the difference between the results of treatment and control group for the level of MAPE reduction in percentage points, we conduct a non-parametric Wilcoxon-Test as we cannot assume a normal distribution and we have two independent groups. Here we cannot use the Fisher's exact test as our target variable is numeric and not categorical as before.

5. Results

First, results are presented per assumption. Second, we summarize and discuss the results in an aggregated fashion and relate them to the key assumption.

A1 (MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction):

In the treatment group, the relative frequency of MPE changing in the right direction after feedback is 91.2%. For the control group the corresponding value (after the blank page) is 65%. This strongly hints toward a consideration of the feedback leading to a systematic adaptation of the judgments that resulted in respective changes of the MPE observed afterwards: if a subject received a negative MPE as feedback, she or he typically gave higher responses to the following questions and vice versa.

The p-value of the Fisher's exact test is 0.0071, thus the result is highly significant at a 1% significance level.

A2 (MPE-feedback induces emphasized adaptation of judgment in the category with the highest MPE, resulting in larger MPE change in this category):

The percentage of MPE changes in the right direction in the category in which the MPE was the highest in the first sequence is 76.5% for the treatment group after the feedback and 50% in the control group after the blank page. The results for A2 hence provide underpinning that subjects selectively make MPE changes and support the presumption that subjects are aware of their category-specific estimation capability and use the feedback in those categories in which they assume their performance to be low, i.e. those with an emphasized mean bias.

For the results of A2, the p-value of the Fisher's exact test is 0.017, which indicates significance of the difference of results between treatment and control group at a 5% significance level.

A3 (MPE-feedback leads to higher MAPE reduction compared to no feedback given):

In the treatment group, 67.6% of the subjects reduced their MAPE after the feedback, compared to 50% of the control group after the blank page (no feedback). This result indicates that subjects seem to reflect on the feedback and use it to change their judgmental behavior in a way that their MAPE decreased after the feedback, in contrast to the control group that did not receive feedback and did not reduce their MAPE on average.

For these results the p-value of the Fisher's exact test is 0.097, thus the results are significant at a 10% significance level.

A4 (MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category):

In the treatment group, 58.8% of subjects made the highest MAPE improvement after the feedback in the category in which the MAPE was the highest in the first sequence, compared to 47.5% of the control group subjects after the blank page (no feedback). Furthermore, for 83.3% of those subjects in the treatment group, where the categories of highest MAPE in sequence one and highest MAPE improvement matched, the total MAPE considering all categories was improved after the feedback. This speaks for a wise and selective usage of the feedback, leading to increased accuracy of estimations.

The p-value of the Fisher's exact test here is 0.23, therefore the difference in the results between treatment and control group are not significant at a 10% significance level. However, due to the small sample sizes, the power of the test is obviously low, and larger sample sizes are required to achieve significant results here.

A5 (MPE-error-feedback leads to higher MAPE reduction particularly in categories with high MAPE in the first sequence compared to auto-correction):

In 67.6% of cases in the treatment group the feedback lead to MAPE improvements, whereas in 57.5% the auto-correction lead to MAPE improvements in the control group. For these results the p-value of the Fisher's exact test is 0.26, indicating that the results are not significant at a 10% level.

We find an average MAPE reduction of 12.48 percentage points in the category with the highest MAPE in the first sequence after the feedback in the treatment group compared to 3.45 percentage points after the blank page when applying auto-correction in the control group. The p-value for the Wilcoxon test is 0.092, for which reason the difference of the results is significant at a 10% significance level.

Although the results are not highly significant, they indicate subjects' capability of reducing the highest errors better (or more) by applying the feedback compared to a non-selective auto-correction.

Overall, we find indication that the group receiving error-feedback considers it selectively to improve judgments and judgmental accuracy. This is the case for overall error reduction as well as for category specific application of the feedback. In addition, the comparison with auto-correction already supports the assumption that large false-corrections with auto-debiasing can be mitigated with an error-feedback approach as proposed.

6. Discussion, Conclusion, and Outlook

The results provide strong support for our key assumption of wise consideration of feedback related to one's own error-pattern.

In particular, the high degree of matches between categories of highest MPE and correct MPE changes as well as MAPE and MAPE improvement represents the capability of humans to recognize error patterns or structures and being able to selectively adapt judgmental behavior accordingly. Reviewing the motivation of this paper, the category matches contribute to the aim of reducing strong false-corrections as errors are decreased the most where necessity for error reduction is the highest. This demonstrates that such a combination of the machine's and human's strengths – the computation and feedback of the MPE through the machine and the usage of the feedback by the human – achieves collaborative intelligence and is a promising direction of future research.

Considering the comparison of feedback versus auto-correction, we can hypothesize that humans applying feedback based on their own error compared to statistical models blindly applying learned error patterns can reduce large false-corrections. This relates to the research by Blanc and Setzer (2015a) who recommend to feed-back the supposed bias to the expert based on estimated model parameters to improve accuracy. Furthermore, it concerns their future research proposition to show experts bias-related feedback of their past forecasts and the forecast of a statistical model and give the expert the opportunity to act upon the feedback to reduce strong false-corrections. In our experiment, we obtained results indicating that this is supported by providing respective feedback to experts.

Our research has the limitation that, due to the COVID-19 pandemic, it has been challenging to conduct experiments with larger numbers of subjects, to be done in presence and not possible online as of the risk of subjects using search engines.

Regarding our future research plan, additional to running more experiments to further support our assumptions with the scenario used in our first experiment, Figure 5 shows further scenarios that will be considered, and how the first scenario is embedded.

The first scenario (X1), the one considered in this article's experiment, considers situations with low complexity for the human and high complexity for the machine. Therefore, in our first experiment the latent topics (here categories) can be considered to be easily detectable by humans, while the machine is unaware of the categories, can only provide aggregated feedback and is also merely able to auto-correct future estimations uniformly. The resulting assumption for X1 is that humans know when to integrate error feedback into their subsequent estimates as they know in which categories they are biased and might perform better when considering the feedback.

The second scenario (X2) considers situations with low complexity for both human and machine, i.e., here the latent topics are known by the machine and, for example, category-specific auto-correction can be applied. Due to human biases that might also be category-specific, the auto-correction performance of the machine might benefit from this information. An option for X2 would be giving feedback for each question with category awareness, in which case it might be more appropriate for the human to generally follow the machine feedback (X2a).

The third scenario (X3) covers situations with high complexity for both human and machine. Experiments with this scenario will contain questions that cannot be clearly assigned to a category. Furthermore, the categories will be latent in nature and the questions will have a rather vague reference to each other so that categories are not obvious. Here, it will be challenging for a machine to provide specific feedback, while the assumption is that the human might still be able to apply the general feedback wisely and selectively based on her or his domain knowledge and latent categories she or he has in mind.

Overall, the intention of the scenarios and our research is to better understand the situations in which feedback of what type can be expected to be beneficial, shedding light on the applicability of the approach in real-world settings.

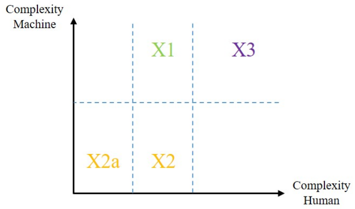


Figure 5. Experimental Scenarios Considered

References

Balzer, W. K., Doherty, M. E., & O'Connor, R. J. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, (106(3)), 410–433.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, (8), 559–573.

Blanc, S., & Setzer, T. (2015a). Analytical debiasing of corporate cash flow forecasts. *European Journal of Operational Research*, (243(3)), 1004–1015.

Blanc, S., & Setzer, T. (2015b). Improving forecast accuracy by guided manual overwrite in forecast debiasing. *Twenty-Third European Conference on Information Systems (ECIS)*, Paper 66.

Blanc, S., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, (69), 3951–3962.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, (36(8)), 887–899.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, (15), 227–246.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, (16), 85–99.

Grant, A., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality*, (30(8)), 821–836.

Jacoby, J., Mazursky, D., Troutman, T., & Kuss, A. (1984). When feedback is ignored: Disutility of outcome feedback. *Journal of Applied Psychology*, (69(3)), 531–545.

Klassen, R. D., & Flores, B. E. (2001). Forecasting practices of canadian firms: Survey results and comparisons. *International Journal of Production Economics*, (70), 163–174.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*, (22), 493–518.

Lawrence, M., & O'Connor, M. (1993). Scale, variability, and the calibration of judgmental prediction intervals. *Organizational Behavior and Human Decision Processes*, (56), 441–458.

Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, (122), 151–160.

Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – a review of the literature. *European Journal of Operational Research*, (213(3)), 459–469.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems*, (16(4)), 339–357.

McCarthy, T. M., Golicic, S. L., & Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting*, (25), 303–324.

Nagar, Y., & Malone, T. (2011). Making business predictions by combining human and machine intelligence in prediction markets. *ICIS 2011 Proceedings*, Paper 20.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, (66(1)), 22–30.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega*, (31(6)), 511–522.

Sargeant, J. M., Mann, K. V., van der Vleuten, C. P., & Metsemakers, J. F. (2009). Reflection: A link between receiving and using assessment feedback. *Advances in Health Sciences Education*, (14), 399–410.

Sasse-Werhahn, L. F., Bachmann, C., & Habisch, A. (2020). Managing tensions in corporate sustainability through a practical wisdom lens. *Journal of Business Ethics*, (163), 53–66.

Sengupta, K. (1995). Cognitive feedback in environments characterized by irrelevant information. *Omega*, (23(2)), 125–143.

Sengupta, K., & Abdel-Hamid, T. K. (1993). Alternative conceptions of feedback in dynamic decision environments: An experimental investigation. *Management Science*, (39(4)), 411–428.

Zellner, M., Abbas, A. E., Budescu, D. V., & A., G. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*.

A Decision Support System Including Feedback to Sensitize for Certainty Interval Size

Nathalie Balla[1]

Catholic University Eichstätt-Ingolstadt, Ingolstadt 85049, Germany
{NBalla@ku.de}

Abstract. In decision-making overconfidence and estimation biases can lead to sub-optimal outcomes and accuracy loss. A debiasing strategy presented in this work is to use feedback based on the error pattern of own previous absolute and 90% certainty (confidence) interval estimates. This is comprised in a decision support system (DSS) and applied in an experiment, where results indicate support for the key assumption that subjects are able to selectively reduce their overconfidence and their estimation bias, if present, with the help of the provided feedback.

Keywords: DSS, interval estimation, overconfidence, overprecision, debiasing

1 Introduction

A key discussion in behavioral operations and information systems research is how to reduce overconfidence and other biases in order to achieve higher quality judgments and decisions. Frequently, experts, for example in supply management, have a sense of control that can lead to too optimistic estimations and they often cannot assess their own performance accurately [1]. Research has found that experts are often overconfident in their judgments and that this leads to low performance and detrimental judgments [2], [1], [3]. Moore and Healy [4] differentiate between three types of overconfidence: overestimation, overplacement, and overprecision. Overprecision denotes being too certain that one's estimate is more accurate than it actually is, to which this work refers to. Ren and Croson [5] investigated the newsvendor problem in inventory decisions and found overprecision leading to underestimation of the variance of demand. They measured overprecision by letting subjects answer ten general knowledge questions each with a 90% confidence interval indicating their certainty of 90% that the true answer lies within this range. In case the true answer lies within the interval in nine out of ten questions, the subject is considered well calibrated. A similar approach based on 90% certainty (confidence) intervals to measure overprecision is used in this work. In decision analysis, estimations regarding intervals are frequently used and Soll and Klayman [6] claim that decision makers' 90% intervals often include the true answer in under 50% of cases, which means they are too sure of themselves and not able to assess their own performance.

International Conference on Operations Research
September 2022, Karlsruhe, Germany

For this reason it is worthwhile to concentrate on interval estimation and its improvement. The method of application of certainty intervals used here corresponds to a typical method identified by Klayman [2]: asking subjects for a single numerical estimate and a 90% confidence interval, that is an upper and lower bound, for which there is a 90% probability that the correct answer lies between them.

In this short paper a novel design for a Decision Support System (DSS) is presented and applied, which aims at reducing overconfidence, particularly overprecision, and estimation biases (under- or overestimation, here meaning systematic too high or too low estimations). According to Ancarani, Di Mauro, and D'Urso [1], benchmarks should be available to experts to let them evaluate their relative performance and a decrease in overconfidence can be achieved by feedback regarding past judgments given in a timely manner. Following this notion, the subjects are provided with feedback regarding their mean percentage error (MPE) on previous absolute answers as well as feedback in which questions the correct answer lay inside the 90% certainty interval. The MPE feedback should give subjects an indication of their relative performance. Two aspects are mainly examined: the average broadening of the intervals before and after the feedback, presumably resulting from the feedback per question if the correct answer lay within the given interval; and the shift of the intervals before and after the feedback, presumably resulting from the MPE feedback, as the MPE points in a certain direction. The assumption is that if the average interval was too narrow to include the correct answer before the feedback, subjects broaden their average interval after the feedback, indicating a reduction of overprecision. The assumption regarding the shifts is that if the average interval was too far away to include the correct answer before the feedback, subjects shift their average interval in direction of the MPE feedback, indicating a reduction of estimation bias. Additionally, the questions come from different categories and it is examined if subjects are able to selectively apply the feedback to intervals in categories in which they performed poorly. The assumption here is that subjects are able to selectively reduce overprecision as well as estimation bias, if present, with the help of the MPE and interval feedback provided. In the following, the experimental design and the assumptions are described. Subsequently, the results are presented and lastly discussed including a conclusion and an outlook.

2 Experimental Design

41 business students took part in the experiment (20 in treatment, 21 in control group), of which 21 are female and 20 male. Subjects were requested to answer point estimate questions and 90% certainty intervals from general knowledge categories: *number of residents of a country*, *river length*, and *mountain height*. Example questions are: "How many residents does France have?", "How long is the Hudson River (in km)?", "How high is the Mount Everest (in meters)?". The experiment is composed of two sequences of each 15 questions, 30 questions overall. The mentioned categories are not communicated to subjects but easy

to detect. All questions include visual cues for estimation support: a map of the respective country with the ten largest cities and an indication of a range of their size; a map of the river including a scale in the legend; a topographical map of the respective mountain and a reference mountain height. These cues are meant to reduce error variance. Subjects are randomly assigned to treatment or control group and the experiment starts with a welcome page and then requires subjects to answer estimation questions and to specify a 90% certainty interval by entering a lower and an upper bound. This certainty interval indicates the range of which subjects are 90% certain that the correct answer lies within it. As soon as the subject entered her answer, the next question is shown. After 15 questions either feedback is provided for 30 seconds (treatment group) or a blank page inviting to a 30 second break is shown (control group). For the feedback, a subject's MPE is computed and shown together with the correct answers. Additionally, the subject sees per question if the correct answer lay within her given interval. The MPE is calculated by taking the difference between the estimate and the actual (actually correct answer) per estimate, dividing this difference by the actual and multiplying it by 100. MPE is therefore the mean of these values. The MPE was selected as measure due to comprehensibility and simplicity to apply it for debiasing. For instance, a negative MPE would mean that the subject needs to place the estimation higher and to shift the certainty interval upwards. MAPE, the mean absolute percentage error, is used as performance criterion to determine improvement or deterioration between question sequences. Because the payouts depend on performance, the MAPE is also used to determine the winner subjects. In the briefing, subjects are given information on how to interpret and apply MPE for debiasing (without telling them they will receive feedback) together with information on MAPE used as performance measure. Moreover, subjects are told the meaning of the 90% certainty intervals. Following the feedback or the blank page, a subject is faced with another sequence of 15 judgments. At the end of the experiment, subjects receive performance information, are debriefed, and their MAPE is calculated. Every subject receives a payout for participation and can additionally win one of two prizes per group. If the MAPE of a subject is lower, her chance is higher to win a prize, which is ought to incentivize subjects.

3 Assumptions Studied

The key assumption is that subjects are able to selectively reduce their overprecision and estimation bias, if present, with the help of MPE and interval feedback. This key assumption is divided into four sub-assumptions A1–A4.

A1: The certainty intervals become broader after the feedback, if they were too narrow to include the correct answer before the feedback, more often in the treatment than in the control group. The presumption is that overprecision exists if subjects specify an interval that is too narrow to include the correct answer. A1 aims to investigate if overprecision can be reduced by, most likely, the feedback on the intervals of the questions. To test A1, the relative frequency of correct answers within the given intervals before the feedback is computed as well as the average relative size difference (between sequence one and two) of the intervals per user. This is done by dividing the average interval size in the second sequence by the one in the first sequence per category and then taking the mean over categories. It is assumed that if the correct answers lay in the given interval in less than 50% of a subject, the intervals are too narrow on average. Results for thresholds of under 33% and 25% of correct answers lying inside the interval are also presented to demonstrate stability of results. If this is the case, the average interval after the feedback should be broadened as the subjects should realize that their intervals were too narrow and they may have been overprecise. For the control group the expectation is that around 50% of subjects make their average interval broader in the second sequence, if they did not include the correct answer in the first sequence (baseline in case of randomness).

A2: The certainty intervals are shifted in the direction of the MPE feedback after the feedback more often in the treatment than in the control group. To test A2, it is analysed how the per category calculated average upper and lower bounds of the intervals are placed before and after the feedback per user. If both of the average bounds moved in the direction of the MPE feedback, that is if the MPE feedback is negative, the bounds should be larger than before the feedback and vice versa, then it can be assumed that the subject reduced her estimation bias by shifting her interval. Ratios are computed per category and treatment group from which the mean is taken over the categories per treatment group.

A3: The certainty intervals become broader more often after the feedback especially in those categories, in which the intervals were too narrow to include the correct answer before the feedback compared to no feedback given. To test A3, the category with the minimum relative frequency of correct answers within the given intervals before the feedback is determined. Then the average interval size per category and the difference of average interval size between sequence one and two per category is calculated by dividing the average interval per category of sequence two by that of sequence one. Therefore, we find the category with the maximum broadening of the average interval. The relative frequency of the matches between these two categories (minimum and maximum) indicates how often subjects selectively applied the feedback to specific categories. The objective is to show a difference in frequency of matches between the treatment groups.

A4: The certainty intervals are shifted in the direction of the MPE feedback after the feedback in those categories, in which the absolute MPE was the highest before the feedback more often in the treatment than in the control group. To test A4, the category with the maximum absolute MPE in the first sequence is determined per user and compared to the categories in which the average interval was shifted in the direction of the MPE in the second sequence. Then the relative frequency of matches between the category with the highest absolute MPE and the categories in which the average intervals were shifted in the direction of the MPE is computed. If the percentage value in the treatment group exceeds the one in the control group, selective application of the feedback to specific categories can be assumed.

For all assumptions we conduct a one-sided Fisher's exact test between the results of treatment and control group.

4 Results

In this section, results are presented per assumption.

A1: 19 subjects in each treatment and control group had the correct answer inside their interval in the first sequence in less than 50%. In 70% of these cases in the treatment group the average interval became broader after the feedback compared to 57.1% in the control group. For a 33% threshold for correct answers lying inside the interval, 12 subjects were in the treatment and 10 in the control group. In the treatment group 65% of subjects broadened their certainty interval after the feedback compared to 52.4% in the control group. For a 25% threshold, 9 subjects were in the treatment and 8 in the control group. Here, in the treatment group 60% of subjects broadened their certainty interval compared to 52.4% in the control group. The p-value of the Fisher's exact test is 0.45, meaning the results are not significant at a 10% level, likely due to the small sample size.

A2: In the treatment group 55% of subjects shifted the certainty interval in the direction of the MPE feedback after the feedback compared to 34.9% in the control group. The p-value of the Fisher's exact test is 0.14, which means the results are not significant at a 10% level, likely due to the small sample size.

A3: In 70% of cases the subjects broadened their average interval in that category the most after the feedback in which the least correct answers lay inside the interval in the first sequence for the treatment group. In 47.6% of cases the subjects achieved this in the control group. The p-value of the Fisher's exact test is 0.13, for which reason the results are not significant at a 10% level, again for the same reason.

A4: In 60% of the treatment group the average intervals were shifted in those categories where the absolute MPE was the highest compared to 23.8% in the control group. The p-value of the Fisher's exact test is 0.02, which means the results are significant at a 5% level.

5 Discussion

The shown results, despite the limited sample size, provide support for the key assumption of the combination of the feedback elements aiding subjects to reduce overprecision and estimation biases selectively. Regarding *A1*, it seems that the difference between treatment and control group becomes larger, when the threshold is higher, that is when the interval is set too narrow more often. This is consistent with expectations, because the more often subjects miss the correct answer, the stronger they are able to adapt their intervals after the feedback, whereas the control group does not have this chance. Most likely the broadening of the intervals, indicating a decrease in overprecision, originates from the interval feedback through which subjects reflect on their previously given lower and upper bounds. By realizing they were too certain in setting their bounds and therefore setting them too close to one another, they became less certain through the feedback and the reflection lead to adaptation of their future bounds, that is a decrease in overprecision. Regarding *A2*, the ratio of subjects shifting their interval in the direction of the MPE is higher for the treatment group. This shifting behavior shows that subjects do not only consider the feedback of the correct answer lying in their interval but also the MPE feedback indicating if their interval was generally too low or high. Applying this feedback to their future upper and lower bounds means they also reflect on the general position of their certainty interval independent of its size, which leads to a reduction of estimation bias, either under- or overestimation. As *A3* and *A4* also show higher ratios for the treatment group, subjects seem to be able to selectively apply the feedback to specific categories, where it is most necessary. Reducing overprecision and estimation bias selectively, leads to an overall error reduction.

6 Conclusion and Outlook

This work indicates that subjects are able to reduce their overprecision as well as their estimation bias, in general and for specific categories that performed poorly before the feedback. A selective consideration of feedback shows that subjects are able to recognize novel patterns and use this knowledge effectively. These two focal aspects in turn result in a general accuracy improvement in terms of MAPE reduction. This research has the limitation that, due to the Corona pandemic, it has been challenging to conduct experiments with large numbers of subjects, for which reason the sample size is rather small. However, this work is research in progress and more experiments will be conducted to support the findings. In addition, this experiment can be conducted with different categories and scenarios such as categories less obvious for subjects.

References

1. Ancarani, A., Di Mauro, C., D'Urso, D.: Measuring overconfidence in inventory management decisions. Journal of Purchasing and Supply Management 22(3), 171-180 (2016). doi:10.1016/j.pursup.2016.05.001
2. Klayman, J., Soll, J.B., Gonzalez-Vallejo, C., Barlas, S.: Overconfidence: It Depends on How, What, and Whom You Ask. Organizational Behavior and Human Decision Processes 79(3), 216-247 (1999). doi:10.1006/obhd.1999.2847
3. Shipman, A.S., Mumford, M.D.: When confidence is detrimental: Influence of overconfidence on leadership effectiveness. The Leadership Quarterly 22(4), 649-665 (2011). doi:10.1016/j.leaqua.2011.05.006
4. Moore, D.A., Healy, P.J.: The Trouble With Overconfidence. Psychological Review 115(2), 502–517 (2008). doi:10.1037/0033-295X.115.2.502
5. Ren, Y., Croson, R.: Overconfidence in Newsvendor Orders: An Experimental Study. Management Science 59(11), 2502-2517 (2013). doi:10.1287/mnsc.2013.1715
6. Soll, J.B., Klayman, J.: Overconfidence in Interval Estimates. Journal of Experimental Psychology: Learning, Memory, and Cognition 30(2), 299–314 (2004). doi:10.1037/0278-7393.30.2.299

Debiasing Judgmental Decisions by Providing Individual Error Pattern Feedback

Nathalie Balla, Thomas Setzer

[a]*Katholische Universität Eichstätt Ingolstadt, Auf der Schanz 49, Ingolstadt, 85049, Germany*

Abstract

We present a novel Decision Support System (DSS) that provides experts with feedback on their personal potential bias based on their previous error pattern. As the feedback stems from an expert's own error pattern, it intends to enhance their self-reflection, foster wise consideration of the feedback and mitigate potential biases. Common DSSs, which typically filter and visualize relevant information, also aim to support rational decisions. However, experts using such DSSs generally still exhibit systematically flawed decisions. These systematic errors can be detected and automatically corrected by DSSs, respectively machines, by correcting humans' judgments with a statistical method afterwards. Nevertheless, this often leads to suboptimal outcomes, because the machine also falsely corrects originally proficient judgments as a machine is, unlike a human expert, not conscious of implicit knowledge or possibly unaware of structural breaks. We assume that experts are able to apply the above described feedback systematically and selectively to different decision tasks and to therefore reduce their potential bias and error. To test this assumption, we conduct experiments with the DSS. Therein, subjects provide point estimations as well as certainty intervals and subsequently receive feedback, which is given by a machine that learns the error pattern of the respective subject based on previous answers. After the feedback, subjects answer further questions. Results indicate that subjects reflect on their own error pattern and apply the feedback selectively to further estimations to reduce overall bias and error.

Keywords: Decision Support System, Debiasing, Feedback, Self-Reflection,

1. Introduction

The accuracy of estimations is of significant importance for businesses as they are the foundation for crucial decisions, which ultimately impacts a company's success. Many of these decisions build on judgmental approaches as decision makers are humans with individual attitudes and estimation heuristics (1; 2; 3). Hence, a key research topic in Decision Support System (DSS) research is debiasing, meaning to ameliorate decision outcomes and informing decision makers about potential biases (4).

Despite using supposedly well designed DSSs that typically support by filtering and visualizing relevant information to foster rational decisions, research demonstrates that decisions still come out systematically erroneous, including biases like overconfidence, mean or regression bias, or anchoring (5; 6; 7; 8; 9; 10; 11).

As an example, the work by Blanc and Setzer (12) motivates the necessity for a novel kind of DSS to mitigate biases. They identify mean and regression biases in expert cash flow forecasts despite the usage of DSSs. The authors find that error patterns can be detected statistically and then corrected automatically, overall enhancing accuracy. Automated correction refers to the correction of experts' forecasts by a statistical model learned on previous experts' errors. However, this auto-correction is applied to expert judgments without consideration of possible differences in certainty in the original judgment. Therefore, well made expert estimations may be corrected in the wrong direction, leading to an unnecessary higher error rate (13).

While this result and the outcomes of other studies show that machines are able to detect consistent judgmental error patterns in practical settings, previous research has also shown that it is beneficial to include humans and machines into more interactive processes to combine their complementary strengths (14; 15). Humans are better at recognizing and detecting novel or transferable situations, or unseen developments based on domain knowledge and intuition and integrating contextual information. Machines are stronger in extracting regular patterns from data (16; 17; 18; 19).

As judgmental error data are also data with potential structures, we present a new DSS that learns patterns in individual error data of an expert's past judgments and feeds back systematic patterns (and potential underlying biases) found to that expert, who then decides how to incorporate the feedback when making further judgments. The intention of the DSS is to increase estimation accuracy and make aware of – and allow to – reduce bias, particularly over- and underestimation and overconfidence. In addition, the DSS intends to mitigate the false-correction problem with auto-correction.

Following the assumption that experts are capable of applying the feedback in a beneficial fashion, the feedback should be rejected if a decision belongs to a domain in which the expert is well versed, or is certain that the feedback is not applicable to the particular judgment at hand. If an expert is less certain to be unbiased, the feedback should be considered.

To test whether this holds true, we conduct experiments in which subjects provide estimations and certainty intervals from varying categories, which are not disclosed. The categories are meant to represent a new structure for the human to recognize, also featuring different expert tasks, in which experts have different levels of knowledge. The feedback consists of the personal mean bias and is provided after a first sequence of questions. It is meant to foster awareness of a personal potential bias, encourage self-reflection on prior errors and contemplation on how to apply the feedback. Results indicate that humans are capable of recognizing new structures and reflecting on own error patterns to systematically and selectively apply feedback to reduce error and bias.

Overall, the contribution of this work comprises a promising approach to achieve collaborative intelligence by using a DSS that involves individual error pattern feedback learned by a machine. We test whether humans are capable of recognizing specific categories, reflecting on own error patterns and wisely and selectively applying feedback to reduce error and bias. Further, we test if humans perform better applying own error feedback compared to machines applying auto-correction. To our knowledge this has not be examined so far.

The paper is structured as follows. In Section 2, previous research on auto-correction, feedback and self-reflection is reviewed. In Section 3, the infrastructure and procedure of our DSS is described. In Section 4, the research design of the first experiment is presented. In Section 5 the results are presented, which are then discussed in Section 6. Finally, in Section 7 the work is summarized and an outlook for future research is provided.

2. Prior Work on Auto-Correction, Feedback, and Self-Reflection

In this section, we review literature on auto-correction, feedback that is related to humans' biases, and self-reflection.

2.1. Auto-Correction

Although the intention of DSSs is to enhance decision making and reduce biases, often DSS-based decisions still exhibit systematic errors. However, DSSs can also be used to detect these systematic errors (whether originally supported by DSSs or not), which is considered for instance by Blanc and Setzer (13). The authors study accuracy gains when applying auto-debiasing. They observe overall accuracy gains with auto-correction, i.e., when replacing the expert predictions with the corrected predictions.

This in line with the findings of Goodwin (20), who discovers mean and regression biases in judgmental sales forecasts and applies statistical correction leading to high cost savings.

However, the results of Blanc and Setzer (13) also show that on their empirical data set the variance of error differences between the original and the corrected forecast increases with the magnitude of a correction; the decile bins with the highest discrepancy have the strongest accuracy improvements and deteriorations. The authors reason that auto-correcting estimates without considering how sure the experts are, can consequently also lead to correcting initially accurate expert estimations in a detrimental way, resulting in large errors especially in outer decile bins (13).

Based on these findings, Blanc and Setzer (12) propose a DSS presenting the expert, after the submission of their judgment, the prediction of a statistical correction model including a specification of the bias possibly pushing the gap between correct estimation and the expert's expectation. The suggestion is to derive this kind of benchmark prediction by correcting error patterns from previous estimations that are consistent over time and inviting the expert to either adopt or to edit the statistical model prediction. In the ideal case, the expert would overwrite those model predictions that lead to large false-corrections (13).

2.2. Feedback

Clearly, the form of feedback plays a primary role in whether and how experts are able to accept or reject the feedback and choosing the right feedback-type is therefore key to a beneficial feedback system.

A known differentiation of feedback types is outcome feedback (OFB) and cognitive feedback (CFB). OFB refers to "information that describes the accuracy or correctness of the response" (21), and constitutes often only the correct answer. CFB is "information regarding the how and why that underlies this accuracy" (21).

Many researchers (22; 23; 6) present evidence that OFB, when simply giving correct answers as feedback, is fairly useless, corresponding to many studies generally considering OFB rather effectless and even obstructive to learning when conscious of the correct result (23). This is due to the lack of informative content within this kind of feedback. However, OFB in other forms can be beneficial. For example, as individual performance feedback, which is shown by Benson and Önkal (24) investigating the latter in probability estimation. The subjects in their experiment make four weekly predictions of football games for the following weekend for the winning probability of a certain team. In the treatment group performance feedback is provided, whereas this is not done for the control group. The results of the experiment show that the performance feedback leads to forecasting accuracy improvement.

Examining OFB and CFB, Sengupta (25) performs experiments, in which participants are prompted to make decisions on personnel screening. OFB together with CFB is provided to the treatment group and only OFB to the control group. In this case, OFB is represented by rating decisions made by the expert committee and CFB by the committee's decision strategy referring to similar jobs, consistency scores, and information regarding a participant's own strategy. Confirming findings of similar research, results illustrate that combining OFB and CFB supports participants in exceeding the performance of those participants only receiving OFB.

As of its particular importance to our work and experiment, we will now review feedback approaches aimed at reducing bias in interval estimation, more specifically, overprecision, which besides overestimation and overplacement is one of the three types of overconfidence. Overprecision refers to being too certain that one's estimate is more accurate than it actually is (26), and awareness of overprecision is therefore key to applying the feedback.

Klayman, Soll, Gonzalez-Vallejo, and Barlas (27) introduce a common method for interval estimation. Participants must give a numerical estimate as well as a 90% confidence interval, referring to an upper and lower bound where the probability is 90% that the true answer lies inside the interval.

Soll and Klayman (28) suggest the presence of an overprecision bias if the true answer lies inside 90% intervals given by a decision makers in less then 50% of judgments. This signifies that a person is overly self-assured of their judgmental accuracy.

Regarding feedback to reduce overprecision, it is recommended that benchmarks should be accessible to an expert to facilitate the evaluation of their relative performance. In this spirit, a decline of overprecision can be accomplished by feedback on previous judgments provided shortly after the first judgments made (29).

Despite not finding research focusing on the selective application of feedback as it is the subject in our setting, the aforementioned results concerning feedback on performance and corrective capability demonstrate their potentials to improve judgments, which strongly motivates the usage of feedback based on error patterns for our purposes.

2.3. Self-Reflection

A prerequisite for the effectiveness of any feedback-DSS is the willingness of acceptance and therefore adoption of feedback by experts, which are often overconfident and neglect recommendations and feedback. This frequently holds true despite being told the opposite by a software, which may be due to the fact that it is external feedback (9).

To overcome this challenge, facilitation of a self-reflective process, that is the interpretation and assessment of own thoughts, emotions, and actions, is necessary (30; 31).

Research related to reflection on feedback by experts is conducted by Goodwin (32) with an experiment in which forecasters are asked to review their judgmental predictions. Requiring forecasters to self-reflect by giving a statement for why they adjust the prediction the way they do, leads to higher performance and stronger accuracy improvement compared to not requiring the latter. Moreover, Sargeant, Mann, van der Vleuten, and Metsemakers (33) conducted interviews with physicians who evaluated assessment feedback they received, showing that reflection is valuable referring to the manner of feedback application.

Overall, prior research suggests to counteract the false-correction issue by feeding back a specification of error patterns and bias learned from previous own error patterns. Regarding the type of feedback, the literature suggests the combination of personalized and performance related OFB with CFB, where it is advisable to also use this type of feedback to reduce overprecision and therewith fostering a wise and selective feedback acceptance and incorporation. Furthermore, mirroring an individual bias the expert is able to comprehend and correct fosters self-reflection, learning, and wisely applied debiasing.

We now propose the anatomy of a novel DSS aimed at collaborative intelligence considering, or operationalizing, these conclusions and fostering an expert's differential confidence in different decision or judgmental tasks. Then, we will present the design of a first experiment to study the efficacy of such types of DSSs.

3. DSS Infrastructure and Procedure

We now describe the infrastructure and procedure of the DSS we propose. The steps followed with the DSS are, from a procedural perspective, depicted in Figure 1.

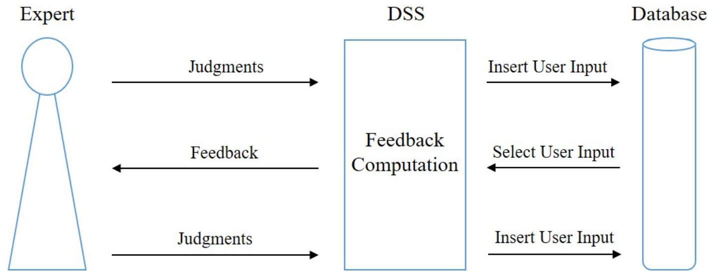


Figure 1: Infrastructure and Processes of DSS

First, an expert's judgments, which might relate to different tasks, or question categories, are captured and stored in the database.

Based on judgments given by an expert and the observed errors, the DSS computes and provides feedback to that expert. This feedback can be of any type, depending on the kind of bias that is aimed to be detected and mitigated, but must be based on the expert's personal prior judgment errors. As an example, in our first experiment, meant to be a basic implementation of the DSS for a proof of concept, the personal error pattern of the subject is represented by the mean percentage error (MPE) computed over their past judgments. We note that this exemplary statistical model can be replaced by other models later on to detect other and more complex biases.

After the feedback, the experts make novel judgments, i.e., they answer novel questions that may relate to the same tasks or categories for questions.

The feedback intends to make aware and confront the expert with their own personal potential error pattern to stimulate self-reflection on previously given judgments and foster thoughtfulness on whether and how to apply the feedback on the next judgments. After potentially recognizing certain categories or domains the questions might relate to, the expert may, for example, also apply the feedback differently to different tasks or categories although no tasks or categories are communicated. Thus, an expert may make use of rather generic feedback, computed over all of their past judgments, in a cognitively wise manner to make novel judgments. Specifically, an expert may have tasks or a categorization in mind in which they would group the questions into, and may have category-specific confidence in this judgmental ability, leading to incorporating the feedback for certain questions and neglecting the feedback for others. Here, it is important to note that an expert may have a personal categorization scheme in mind, which may be very different to the one of a statistical model, motivating the mirroring of more aggregated feedback. An example will be provided in the next section where the experimental design is presented.

After another set of judgments have been captured by an expert, feedback is again computed and mirrored to the expert to allow for continuously adjusting debiasing and correction over time.

Regarding the technical part, the DSS has been developed using Dynamic HTML (PHP) as frontend, a Relational Database Management Server (MySQL) as backend storing different parameters for the DSS, questions and answers given by experts. The DSS is designed as a Web-Application so it can be easily used on every computer. The tools used for determining error patterns, displaying them as feedback, and computing associated loss functions are developed in PHP and *R*.

In this section, we introduced the general procedure, infrastructure and intention of the DSS, and will now describe the actual implementation of the DSS for our first experiment. To show how the scenario considered in our first experiment fits into the larger picture of our research agenda, further scenarios for future research are sketched in Section 7.

4. Experimental Research Design and Hypotheses

First, the research design of the experiment is described, whereby information thereon has been already provided in the previous section to make the idea and concept of the DSS more tangible. Second, the hypotheses and corresponding measures for analysis are presented.

4.1. Research Design

Our gender-balanced sample of subjects consists of 97 students from different fields between 18 and 31 years old. They are randomly assigned to treatment (51 subjects) and control group (46 subjects).

The configuration of the experiment is stored in a database. This includes the estimation questions, the corresponding correct answers and visual cues, the form and timing of feedback, pages for briefing and debriefing, comprehension questions, rules when an experiment terminates, and a final questionnaire.

After comprehension questions are answered, subjects are required to answer point estimate questions from general knowledge categories. In addition, subjects are asked to indicate a 90% certainty interval for every question. The categories contain questions about *number of residents of a country*, *river length*, and *mountain height*. Example questions are: "How many residents does France have?", "How long is the Hudson River (in km)?", "How high is the Mount Everest (in meters)?". Categories are neither communicated to subjects nor used by the correction model, but arguable easy to anticipate by humans.

This scenario intends to simulate expert judgment by supposing experts have expertise and basic confidence in all domains they are responsible for, similar to the subjects in the experiment who most likely have basic knowledge of the general knowledge questions. The categories in the experiment are meant to mimic different domains, where experts as well as the subjects may apply different heuristics and perform better in some than in others. Due to these categories or types of questions, wherein proneness to biases is proposed, humans may recognize patterns that a machine would not be able to detect. Subjects may also have a different categorization in mind such as regions or continents of the world, in case a subject realizes they is more adept at questions regarding, for example, Europe compared to Asia.

Having the named categories in mind, a mountaineer will have great knowledge of mountains and will probably make quite accurate estimations in this category even before the feedback and being self-aware of this knowledge, may not apply the feedback or may apply it less pronounced to subsequent mountain height questions compared to questions in other categories.

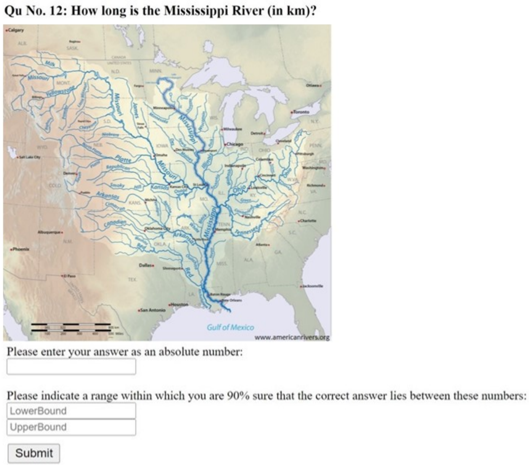
With every question a visual cue for estimation support and to reduce error variance is displayed. This again mimics the environment of experts, which are also supported by orientational data, figures or graphs when making a decision.

For estimations of residents in a country, a map of the respective country including the ten largest cities with an indication of a range of their size is presented. For river lengths, a map of the respective river with a scale in the legend and for mountain heights, a topographical map with a reference mountain height is shown. An exemplary estimation question is depicted in Figure 2. In this example, the subject is prompted to provide an estimation for the number of kilometers of the Mississippi River as well as a range within which they is 90% sure that the correct answer lies inside by indicating an upper and lower bound. For estimation support, a map of the Mississippi including a scale is displayed with the question.

In total, there are 30 questions divided in two sequences with an interruption after 15 questions of minimum 30 seconds. During the interruption, the treatment group is confronted with feedback and the control group receives a blank page inviting to take a break. After the interruption both groups receive another 15 new and unseen questions from the same categories. Figure 3 shows an exemplary feedback page with the mean bias of the subject.

The feedback comprises a subject's own MPE computed across its first 15 point estimates as well as information per question on its given answer, the actually correct answer, and if its given interval includes the correct answer.

The MPE is computed by taking the difference between the given answer



Qu No. 12: How long is the Mississippi River (in km)?

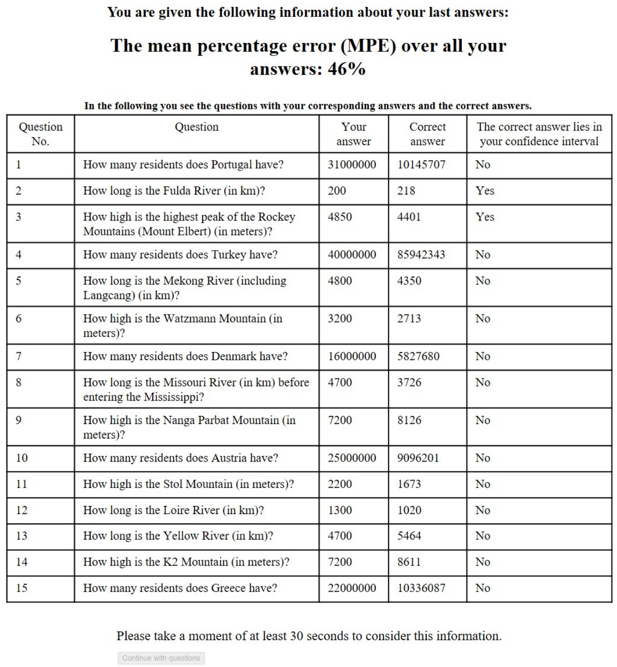
Please enter your answer as an absolute number:

Please indicate a range within which you are 90% sure that the correct answer lies between these numbers:

Figure 2: DSS Interface – User Prompt Example

and correct answer per question, dividing this difference by the respective correct answer, multiplying it by 100, and taking the mean of this over all the previous answers and thus over all categories. We choose the MPE for the first experiment as it is a comprehensible statistic and theoretically simple in application even though it is not trivial for subjects to apply the feedback correctly to unseen questions. For instance, an MPE of 50% means that given answers exceed correct answers by 50% on average and application thereof would mean to take $\frac{2}{3}$ of an upcoming estimate. The information on the individual answers give an indication for which categories push the MPE, or where over- or underestimation is discovered to encourage reflection on the manner of adapting future estimations. As the feedback on intervals is meant to give guidance if their intervals were set too narrow, possibly due to being overconfident in their answer, this may lead to decreasing overconfidence and higher self-reflection. We note that feedback is strictly related to patterns in a subject's own error history.

For performance determination we calculate the mean absolute percentage error (MAPE) per subject, which is similar to the MPE computation with the difference that absolute values of the percentage errors are averaged so that



You are given the following information about your last answers:

The mean percentage error (MPE) over all your answers: 46%

In the following you see the questions with your corresponding answers and the correct answers.

Question No.	Question	Your answer	Correct answer	The correct answer lies in your confidence interval
1	How many residents does Portugal have?	31000000	10145707	No
2	How long is the Fulda River (in km)?	200	218	Yes
3	How high is the highest peak of the Rockey Mountains (Mount Elbert) (in meters)?	4850	4401	Yes
4	How many residents does Turkey have?	40000000	85942343	No
5	How long is the Mekong River (including Langcang) (in km)?	4800	4350	No
6	How high is the Watzmann Mountain (in meters)?	3200	2713	No
7	How many residents does Denmark have?	16000000	5827680	No
8	How long is the Missouri River (in km) before entering the Mississippi?	4700	3726	No
9	How high is the Nanga Parbat Mountain (in meters)?	7200	8126	No
10	How many residents does Austria have?	25000000	9096201	No
11	How high is the Stol Mountain (in meters)?	2200	1673	No
12	How long is the Loire River (in km)?	1300	1020	No
13	How long is the Yellow River (in km)?	4700	5464	No
14	How high is the K2 Mountain (in meters)?	7200	8611	No
15	How many residents does Greece have?	22000000	10336087	No

Please take a moment of at least 30 seconds to consider this information.

Figure 3: DSS Interface – Feedback Page Example

the errors cannot balance each other out as it is with the MPE. Therefore, we can identify if a subject improves or deteriorates in performance before and after the feedback or blank page, which is also important for payouts. In the briefing prior to the experiment, subjects receive information on how the MPE can be interpreted (without telling subjects they will receive feedback) including information on the performance measure and its impact on payouts.

After the second sequence of questions, all subjects receive the above described feedback. The experiment ends with feedback from subjects and demographic questions. Every subject receives a payout for participation and is able to win an additional prize money per group depending on their MAPE-performance. The lower the MAPE, the higher is the probability to win a prize, which is meant to incentivize subjects. The experimental procedure is illustrated in Table 1.

Table 1: Experimental Design and Procedure

	Treatment Group	Control Group
15 Questions	x	x
Feedback	x	-
15 Questions	x	x
Feedback and Demographic Questions	x	x

As described above, the intuition of using question categories is many-fold to imitate experts' working environment. This includes: first, the assumption of a sound general knowledge in their field of expertise with emphasized skills in some subfields; second, different estimation heuristics in different subfields; and third, subjects having category-specific depth of knowledge, judgmental ability, and error levels, fostering a category-specific consideration of the error-feedback.

Subsequent to the conduct of the experiment, we will undertake a 90% Winsorization on the APE values before further analysis of results, to prevent potentially strong outliers distorting the results. We will set the low border as the 5%-quantile and the high border as the 95%-quantile of the APE values.

In the following, we will describe the hypotheses tested in the experiment.

4.2. Hypotheses

We now formulate seven hypotheses (H1-H7) for the key assumption that humans are capable of recognizing new structures and reflecting on own error patterns to selectively apply feedback to reduce error and bias.

H1 and H2 relate to adjustment behavior in the right direction after the feedback. This is examined overall as well as category-specific, meaning if change is emphasized in categories where MAPE is higher before the feedback. H3 and H4 refer to accuracy enhancement after the feedback. This is again analyzed in total and category-specific. H5 specifically concerns auto-correction versus human-correction with feedback. H6 and H7 relate to the certainty interval estimation and the reduction of overprecision by examining if subjects broaden their certainty interval after the feedback, again once in general and once category-specific.

H1: MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction. For H1, the MPE is determined per subject as described in Section 4.1 for the first and second sequence, meaning before and after the feedback or blank page. An example for the meaning of right direction is, if a subject has an MPE of -30% in the first sequence, suggesting underestimation, and an MPE above -30% in the second sequence, this hints to an acceptance and incorporation of the feedback to counteract underestimation.

The expectation is that the ratio of right-direction MPE changes in the treatment group exceeds the ratio in the control group and the control group reaches a ratio around 50% as no feedback is given to possibly cause systematic MPE change. To find significance for H1, a Fisher's exact test of independence between the ratios of right-direction MPE changes in treatment and control group is conducted.

H2: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE, resulting in larger MPE change in this category. To test H2, the MPE per subject and category for the first and second sequence and the MAPE for the first sequence is computed. Then, the category with the highest MAPE in the first sequence as well as the category with the greatest MPE change in the right direction in the second sequence is identified, again per subject.

If these categories between first and second sequence match, it indicates that the subject adjusts their estimations in the right direction the most in that category where it is most necessary. This would indicate that the subjects do not blindly adopt the feedback and apply it to all questions in the second sequence, like a course-grained statistical method would do, but are able to use it selectively.

The ratios of category-matches are compared between treatment and control group, expecting the treatment group to achieve a higher ratio. If the categories match in more than $\frac{1}{3}$ of cases (the baseline in case of randomness) in the treatment group, we can assume a sound category-specific application of the feedback. For H2 we conduct a Fisher's exact test between the ratios of category-matches between treatment and control group to detect the significance of found differences in ratios.

H3: MPE-feedback leads to higher MAPE reduction compared to no feedback. For each subject the MAPE is computed for the first and second sequence of estimations. Then, it is analyzed per subject if an increase or a decrease in MAPE between the sequences can be observed. The ratio of MAPE reductions between treatment and control group are compared, where we expect the treatment group to exhibit more MAPE reductions than the control group. Due to absent feedback and hence random MAPE increases or decreases, we expect a ratio around 50% in the control group. A Fisher's exact test is conducted for H3 to verify a significant difference in the results between treatment and control group.

We note that the difference of H3 to H1 is that H1 solely observes the changes of MPE direction, whereas H3 deals with changes of MAPE as accuracy measure. It is possible that the MPE of a subject changes in the right direction without the MAPE being improved when adjusting the estimation too strong. In this case the feedback is adopted by the subject but applied too intensely.

H4: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category. Per subject and category we compute the MAPE for the first and second sequence. Subsequently, we determine the category with the highest MAPE in the first sequence and the category with the highest MAPE decrease in the second sequence. Then, we count the number of these category-matches between first and second sequence for both groups. If the treatment group achieves a higher ratio of category-matches compared to the control group and the former also reveals category matches in more than $\frac{1}{3}$ of cases, we can assume a general ability of selective application of the feedback. Again, we use the Fisher's exact test to find significance in the difference of results between treatment and control group.

H5: MPE-feedback leads to higher MAPE reduction compared to auto-correction.
For H5 we compare the MAPE improvement between the auto-corrected answers in the second sequence and the answers given by subjects in the second sequence in the treatment group. The hypothetical answers the auto-correction would provide in the second sequence are computed from answers of the control group of the second sequence. This hypothetical answer per question is determined by taking the subject's answer and dividing it by the corresponding MPE+1. This simulates a statistical method that applies the error pattern across all future judgments the subject makes without differentiating between questions. Then, we compute the MAPE of these answers in the second sequence and evaluate per subject if there is an improvement from the first to the second sequence. This ratio of MAPE reductions of auto-corrected judgments in the control group is compared to the human corrected judgments in the treatment group (from H3). We expect the ratio of MAPE reductions in the treatment group to exceed the MAPE reductions by auto-correction in the control group. We also conduct a Fisher's exact test to find significance between human-corrected results and auto-correction.

Additionally, to figure out how auto-correction would perform if the machine would be able to correct the answers category-specifically, we calculate the MPE per category per subject in the control group and correct the answers in the second sequence with the respective category-specific MPE. Of these auto-corrected answers we calculate the MAPE per subject and compare it to the MAPE of the first sequence.

H6: The certainty intervals become broader after the feedback, if they were too narrow to include the correct answer before the feedback, more often in the treatment than in the control group.
Regarding H6, the assumption is that if subjects set their upper and lower bound too narrow to each other so that the correct answer does not lie in-between them, they are overprecise.

To test H6, per subject the relative frequency of where the correct answer lies outside the subject's interval in the first sequence is determined. We set a threshold that if the correct answer lies inside the interval in less than 50% of cases in the first sequence, the average interval is set too narrow. We perform the same analysis for a threshold of 35% to demonstrate robustness of results. Additionally, the average relative size difference between first and second sequence of the intervals is computed. Thereby, we normalize interval size by dividing each difference between upper and lower bound by the point estimate answer given by the subject in order to make the interval size comparable. Then, we divide the average interval size in the second sequence by the one in the first sequence to receive the average interval size difference per subject. Finally, the ratio of cases where the interval is too narrow and is then broadened are compared between the groups. We anticipate the treatment group to exceed the control group and the latter to reach a ratio around 50%. Also for H6 we conduct a Fisher's exact test to find significance for the results.

H7: The certainty intervals become broader more often after the feedback especially in those categories, in which the intervals were too narrow to include the correct answer before the feedback compared to no feedback given.
The same analysis as for H6 is done but per category. Thus, per subject and per category the relative frequency of correct answers lying outside the interval as well as the average relative size difference of the intervals between first and second sequence including normalization of interval size as above is computed. Then, the category with the maximum relative frequency of correct answers lying outside the interval and the category with the maximum average relative size increase of intervals is identified. The ratio of category-matches between the sequences is compared between the groups and expected to be higher for the treatment group. We conduct a Fisher's exact test for significance detection.

In total, these seven hypotheses are meant to investigate the (selective) change in behavioral judgment resulting in accuracy improvement originating from feedback based on personal error patterns. Additionally, they intend to examine whether and how humans can reduce the MAPE stronger than autocorrection.

5. Results

In this section, the results are presented according to the hypotheses. [1]

H1: MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction.
In 82.4% of cases the subjects in the treatment group made MPE changes in the right direction after the feedback. The subjects in the control group made MPE changes in the right direction in 54.3% of cases. The p-value of the Fisher's exact test is 0.0027, which demonstrates significance of the difference between treatment and control group at a 5% significance level.

H2: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE, resulting in larger MPE change in this category.
In the treatment group, 62.7% of the subjects made the greatest MPE change in the right direction in that category where the MAPE was the highest in the first sequence. This was the case for 43.5% in the control group. The p-value for the Fisher's exact test is 0.0447, which means the difference in results is significant at a 5% significance level.

H3: MPE-feedback leads to higher MAPE reduction compared to no feedback.
For 64.7% of subjects in the treatment group there was a MAPE reduction after the feedback compared to 52.2% in the control group after the blank page. For H3 we could not determine a significance at a 5% significance level with a p-value of 0.1479 of the Fisher's exact test.

H4: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category.
64.7% of subjects in the treatment group made the largest MAPE improvement in the second sequence after the feedback in that category with the highest MAPE in the first sequence. This was the case for 52.2% in the control group

[1] Of the 97 subjects in the experiment, seven subjects missed one question (due to technical problems), for which reason one answer is missing for each of these seven subjects. However, they are nonetheless included in the analysis as we assume that the single missing answer will not severely impact results.

after the blank page. The p-value for the Fisher's exact test is 0.1479, for which reason we cannot state significance at a level of 5%.

H5: MPE-feedback leads to higher MAPE reduction compared to auto-correction. Compared to the first sequence of the control group the auto-correction yielded a MAPE improvement in 26.1% of cases in the second sequence. In the treatment group, 64.7% of subjects improved their MAPE after the feedback. A p-value of 0.000132 of the Fisher's exact test demonstrates significance at a 5% significance level for the greater MAPE decline through human correction with feedback than auto-correction.

Regarding the category-specific auto-correction, there was a MAPE improvement compared to the first sequence in 50.0%. This result improved heavily compared to non category-specific auto-correction (26.1%), however, it is still inferior to the human-corrected results with feedback (64.7%).

H6: The certainty intervals become broader after the feedback, if they were too narrow to include the correct answer before the feedback, more often in the treatment than in the control group.
The 50% threshold, meaning that the correct answer lay inside the given interval in under 50%, involves 41 subjects in the treatment group and 35 subjects in the control group. Of these 41 subjects in the treatment group, 68.3% broadened their certainty interval after the feedback in the second sequence. Of the 35 subjects in the control group 34.3% broadened their certainty interval after the blank page in the second sequence. This result is significant to a 5% significance level with a p-value of 0.0030 for the Fisher's exact test. The 35% threshold includes 34 subjects in the treatment and 27 in the control group, where 70.6% of the treatment group broadened their certainty interval vs. 29.6% of the control group. This result is significant shown by the p-value of 0.0016 of the Fisher's exact test.

H7: The certainty intervals become broader more often after the feedback especially in those categories, in which the intervals were too narrow to include the correct answer before the feedback compared to no feedback given.
In the treatment group 60.8% of the subjects broadened their interval after the

feedback in the second sequence in exactly that category in which the correct answers lay outside the interval the most before the feedback in the first sequence. This was the case for 50.0% in the control group. The p-value for the Fisher's exact test is 0.1941 for which reason the difference in results between the groups is not significant at a 5% level.

Summarizing, subjects receiving feedback based on their own error pattern achieve a higher error reduction than subjects without feedback. This holds true in general as well as for the category-specific, selective application. Moreover, the subjects in the treatment group are able to accomplish higher accuracy after the feedback compared to a statistical method using the MPE for auto-correction, which illustrates the mitigation of the false-correction problem. Additionally, we find that subjects receiving feedback indicate a larger decrease in overprecision compared to subjects not receiving feedback, also valid for category-specific application.

6. Discussion

In this section, first we discuss the results in general and second, subjects' answers to the usage of feedback during the experiment.

6.1. Discussion of Results

The results presented above demonstrate support for each of the hypotheses, meaning the treatment group showing stronger performance than the control group or auto-correction, with four out of seven hypotheses being significant to a 5% level, whereas also the non-significant results point into directions supporting the underlying assumptions.

Overall, we observe strong support for the key hypothesis of wise and selective consideration and application of feedback based on one's own error pattern leading to bias reduction and accuracy improvement.

Specifically, we detect a high proportion of subjects in the treatment compared to the control group matching the categories between highest MAPE in the first sequence and strongest MPE change in the right direction but also strongest MAPE reduction in the second sequence. This underpins the ability of humans to detect novel structures and reflect on the own error feedback and use it selectively for further judgments. Therefore, errors are reduced the most where they are the largest, leading to an overall greater error decrease.

The findings of H5 indicate benefits and human skills to use feedback wisely compared to auto-correction. Even when the machine would know the categories, the subjects still performed better, i.e. improved their MAPE more, and the combination of the machine providing feedback and the human applying it leads to a reduction of false-corrections and a promising approach of collaborative intelligence for accuracy improvement.

The results of H6 and H7 indicate that subjects receiving feedback show a lower degree of overprecision in the second sequence than subjects not receiving feedback. This finding also indicates that subjects reflect on their own errors, willing to adapt judgment behavior to reduce overprecision.

An interesting observation is that more subjects in the treatment group were able to adjust their MPE in the right direction (H1: 82.4%) than reduce their MAPE (H3: 64.7%) after the feedback. 21.6% of subjects in the treatment group adjusted their MPE in the right direction without improving their MAPE. One likely explanation is that these subjects adjusted their estimations with the right intention, but too excessively such that their MAPE increased after the feedback. This is confirmed by inspecting the individual answers of the subjects. In 45.5% out of the 21.6% observation, the subjects' errors indicate a strong over- or underestimation in the category of resident numbers before the feedback and then surpass the optimal level of adjustment after the feedback so that other categories where the MAPE has decreased could not balance this out. This raises the question how the feedback could be modified to mitigate the problem of over-adjustment beyond beneficial levels, for instance by sensitizing subjects for the magnitudes of adjustment and the risk of potential over-steering.

To strengthen the statements, we conducted two additional experiments: One experiment with different question categories, and another one with non-randomized experiments per subject, where every subject received the same questions in the same order. The results of both experiments are comparable to the results obtained in the first experiment described in this paper and therefore underline our hypotheses. The details thereof are described in the appendix.

6.2. Subjects' Answers to the Reception and Utilization of the Feedback

The outcomes of the experiment generally mirror the subjects' intentions and thoughts during the experiment. At the end of every experiment, each subject in the treatment group was asked the following question: "During the experiment, how did you use the information of the feedback? If you did not use it, why not?". 25.1% of subjects explicitly stated to have adjusted their estimation in a certain direction due to their recognition of over- or underestimation after the feedback. 30.8% even claim to have adjusted their estimation in a selective, category-specific manner. Furthermore, 34.4% of subjects indicate to have broadened their certainty intervals after the feedback.

These declarations show that and how subjects reflected on the feedback and how they intended to apply it for debiasing. In the following, selected quotes of subjects that were given to the question above are presented.

"Calculating the "measured" length of rivers times 2 instead of rounding up (because they are not straight), setting the confidence intervall wider"

"I saw that especially concerning the resident number, I guessed too high, so I tried to reduce it. Also, I saw that is is better to have a higher answer confidence range so that the correct answer ist in the range."

"I looked how much in general I was off with my numbers. For example I doubled most of my numbers after the feedback in the county category and I added a 0 on the length of all the following river numbers I thought were right."

Hence, we assume that the feedback works effectively and fosters reflection on own error patterns and wise and systematic application on further judgments.

7. Conclusion and Outlook

This article demonstrates that humans are able to recognize novel structures (categories) to reflect on feedback based on own errors, and to use it systematically and selectively to achieve overall bias and error reduction. It also shows that humans can achieve stronger error reduction using the feedback compared to a machine applying auto-correction.

With respect to the big picture of our research plan, the experiment conducted and described in this article considers a first, rather straight forward scenario out of three different scenario types.

Figure 4 visualizes the three scenarios and their relation to one another, where in this work we addressed scenario X1, and scenario X2 and X3 are foreseen for future experiments.

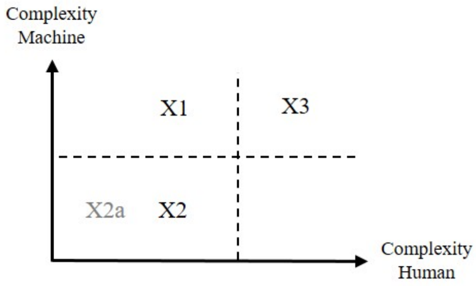


Figure 4: Experimental Scenarios Considered

The scenario X1 examines the condition of low complexity for the human and high complexity for the machine. This condition is designed by the latent topics (categories), which are quite straightforward to identify by humans. Machines were not aware of the categories and could solely offer aggregated feedback and are also only able to auto-correct the following judgments homogeneously. Therefore, the assumption, already tested here, is that humans know in which categories they may be biased leading to a higher performance as they know if and how to incorporate the feedback.

In scenarios of type X2 we will examine the condition of low complexity for human and machine, meaning that more questions will be asked (such that the machine has more data to learn error pattern) and the latent topics are known by both such that auto-correction can be done in a category-specific manner. Because humans' biases may be category-specific, the contribution of the auto-correction by the machine may profit from the information about latent topics. One variation of X2 will be to let the machine provide category-specific feedback for the human for whom it may be reasonable to adhere to this feedback (X2a).

Scenario X3 is meant to study the condition of high complexity for human and machine, where latent categories may be overlapping and questions not distinctly assignable. There should only be a vague connection between the questions such that question types or categories are not evident. It is difficult for a machine to recognize a structure and give precise feedback. In such challenging scenario it will be more demanding to provide feedback to the human that enables them to apply the feedback wisely based on knowledge about the environment by intuition or some latent categories identified.

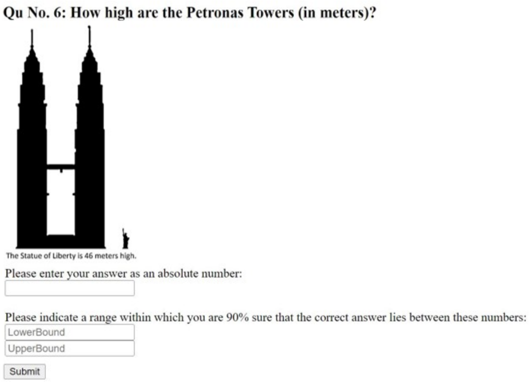
Aside from our described research plan, there are many more opportunities to make use of the proposed DSS concept. An example would be to use probability estimation instead of general knowledge estimation and ask subjects to predict the probability of specific events. Here, the feedback could contain, for instance, the Brier Score to show the subjects how well their estimations are calibrated.

Also, additional measures can be deployed, such as sensors for eye tracking, pulse measurement, or voice when subjects enunciate their thoughts while making decisions. This can help to attain more insight in how humans think about and use machine feedback based on own errors.

In total, this work demonstrates that the investigated debiasing approach has great potential, for which we think we laid a cornerstone to gain further findings. Our research has the purpose to achieve a greater insight in which situations which kind of feedback can be profitable. Therewith, we aim to foster the clarification of the practicability of the concept for real-world business situations.

Appendix A. Additional Experiments

We conducted an additional experiment with different categories, where 32 subjects are randomly assigned to the treatment and 29 to the control group. The categories for this experiment were *beeline distances between cities worldwide*, *number of calories in a certain food*, and *heights of famous buildings*. The visual cues for these categories are respectively a map showing the distance between the two cities without a scale but a hint of the beeline distance between Berlin and Paris, a nutrition table excluding the calories, and a picture of the respective building next to the statue of liberty or a one family house with its height as a reference. An example of the third category is pictured in Figure A.5.



Qu No. 6: How high are the Petronas Towers (in meters)?

The Statue of Liberty is 46 meters high.

Please enter your answer as an absolute number:

Please indicate a range within which you are 90% sure that the correct answer lies between these numbers:

Figure A.5: DSS Interface – User Prompt Example with New Category

We undertook the same analysis for this data and found strong support for the first five hypotheses and rather less support for hypotheses H6 and H7. For H1 90.6% of subjects in the treatment group changed their MPE in the right direction versus 62.1% in the control group with a p-value for the Fisher's exact test of 0.0089, showing significance. Regarding H2, 71.9% of the treatment group made the highest MPE adjustment in the right direction in the second sequence in that category where their MAPE was the highest in the first sequence compared to 48.3% in the control group with a p-value of

0.05215, almost significant to a 5% level. We also found significant results for H3 with 75.0% of subjects in the treatment group decreasing their MAPE after the feedback compared to 44.8% in the control group with a p-value of 0.0156. Referring to H4, 78.1% of the treatment group reduced their MAPE the most in the second sequence in that category where their MAPE was the highest in the first sequence, whereas this was the case for 55.2% of the control group with the difference in results almost being significant with a p-value of 0.0508. For H5 the results were highly significant, where compared to the first sequence of the control group the auto-correction yielded a MAPE improvement in 48.3%. This is compared to 75.0% of the treatment group showing improvement of MAPE after the feedback. These results reveal a high difference with a p-value of 0.02928 for the Fisher's exact test. The results for H6 and H7 are not very supportive and not significant, possibly due to the small sample size.

In total, the results of this second experiment with the new categories underpin H1-H5 of the first experiment and indicate independence of categories, and foster generality of these statements.

In addition to the experiments described above, where we randomized the questions themselves as well as their order for each subject, we also conducted another experiment, in which every subject had the same questions and order of questions. In this non-randomized experiment we used the same categories as in the main experiment and it yielded highly similar results for all of the hypotheses, providing additional support for our hypotheses, as the comparison between treatment and control group is stronger when the order of questions is the same between all subjects in case of small sample size.

References

[1] R.D. Klassen, B.E. Flores, Forecasting practices of Canadian firms: Survey results and comparisons, International Journal of Production Economics. 70 (2001) 163-174. https://doi.org/10.1016/S0925-5273(00)00063-3

[2] T.M. McCarthy, S.L. Golicic, J.T. Mentzer, The Evolution of

Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices, Journal of Forecasting. 25 (2006) 303-324. https://doi.org/10.1002/for.989

[3] N.R. Sanders, K.B. Manrodt, The Efficacy of Using Judgmental Versus Quantitative Forecasting Methods in Practice, Omega. 31 (2003) 511-522. https://doi.org/10.1016/j.omega.2003.08.007

[4] D. Arnott, S. Gao, Behavioral economics for decision support systems researchers, Decision Support Systems. 122 (2019) 113063. https://doi.org/10.1016/j .dss.2019.05.003

[5] S. Blanc, T. Setzer, When to choose the simple average in forecast combination, Journal of Business Research. 69 (2016) 3951-3962. https://doi.org/10.1016/j .jbusres.2016.05.013

[6] M. Lawrence, P. Goodwin, M. O'Connor, D. Önkal, Judgmental forecasting: A review of progress over the last 25years, International Journal of Forecasting. 22 (2006) 493-518. https://doi.org/10.1016/j .ijforecast.2006.03.007

[7] M. Lawrence, M. O'Connor, Scale, Variability, and the Calibration of Judgmental Prediction Intervals, Organizational Behavior and Human Decision Processes. 56 (1993) 441-458. https://doi.org/10.1006/obhd.1993.1063

[8] M. Lawrence, M. O'Connor, B. Edmundson, A field study of sales forecasting accuracy and processes, European Journal of Operational Research. 122 (2000) 151-160. https://doi.org/10.1016/S0377-2217(99)00085-5

[9] J. Leitner, U. Leopold-Wildburger, Experiments on forecasting behavior with several sources of information – A review of the literature, European Journal of Operational Research. 213 (2011) 459-469. https://doi.org/10.1016/j .ejor.2011.01.006

[10] J.S. Lim, M. O'Connor, Judgmental forecasting with interactive forecasting support systems, Decision Support Systems. 16 (1996) 339-357. https://doi.org/10.1016/0167-9236(95)00009-7

[11] J.F. George, K. Duffy, M. Ahuja, Countering the anchoring and adjustment bias with decision support systems, Decision Support Systems. 29 (2000) 195–206. https://doi.org/10.1016/S0167-9236(00)00074-9

[12] S. Blanc, T. Setzer, Analytical Debiasing of Corporate Cash Flow Forecasts, European Journal of Operational Research. 243 (2015) 1004-1015. https://doi.org/10.1016/j .ejor.2014.12.035

[13] S. Blanc, T. Setzer, Improving Forecast Accuracy By Guided Manual Overwrite in Forecast Debiasing, Twenty-Third European Conference on Information Systems (ECIS). 66 (2015).

[14] T. Haesevoets, D. De Cremer, K. Dierckx, A. Van Hiel, Human-machine collaboration in managerial decision making, Computers in Human Behavior. 119 (2021) 106730. https://doi.org/10.1016/j .chb.2021.106730

[15] R. Pinto, T. Mettler, M. Taisch, Managing supplier delivery reliability risk under limited information: Foundations for a human-in-the-loop DSS, Decision Support Systems. 54 (2013) 1076-1084. https://doi.org/10.1016/j .dss.2012.10.033

[16] R.C. Blattberg, S.J. Hoch, Database Models and Managerial Intuition: 50% Model + 50% Manager, Management Science. 36 (1990) 887-899. https://doi.org/10.1287/mnsc.36.8.887

[17] Y. Nagar, T. Malone, Making Business Predictions by Combining Human and Machine Intelligence in Prediction Markets, ICIS 2011 Proceedings. 20 (2011).

[18] M. Arvan, B. Fahimnia, M. Reisi, E. Siemsen, Integrating Human Judgement into Quantitative Forecasting Methods: A Review, Omega. 86 (2019) 237-252. https://doi.org/10.1016/j.omega.2018.07.012

[19] M. Zellner, A.E. Abbas, D.V. Budescu, A. Galstyan, A survey of human judgement and quantitative forecasting methods, Royal Society Open Science. (2021). https://doi.org/10.1098/rsos.201187

[20] P. Goodwin, Statistical Correction of Judgmental Point Forecasts and Decisions, Omega. 24 (1996) 551-559. https://doi.org/10.1016/0305-0483(96)00028-X

[21] J. Jacoby, D. Mazursky, T. Troutman, A. Kuss, When Feedback is Ignored: Disutility of Outcome Feedback, Journal of Applied Psychology. 69 (1984) 531-545. https://doi.org/10.1037/0021-9010.69.3.531

[22] W. Remus, M. O'Connor, K. Griggs, Does Feedback Improve the Accuracy of Recurrent Judgmental Forecasts?, Organizational Behavior and Human Decision Processes. 66 (1996) 22-30. https://doi.org/10.1006/obhd.1996.0035

[23] W.K. Balzer, M.E. Doherty, R.Jr. O'Connor, Effects of Cognitive Feedback on Performance, Psychological Bulletin. 106 (1989) 410-433. https://doi.org/10.1037/0033-2909.106.3.410

[24] P.G. Benson, D. Önkal, The effects of feedback and training on the performance of probability forecasters, International Journal of Forecasting. 8 (1992) 559-573. https://doi.org/10.1016/0169-2070(92)90066-I

[25] K. Sengupta, Cognitive Feedback in Environments Characterized by Irrelevant Information, Omega. 23 (1995) 125-143. https://doi.org/10.1016/0305-0483(94)00061-E

[26] D.A. Moore, P.J. Healy, The Trouble With Overconfidence, Psychological Review. 115 (2008) 502–517. https://doi.org/10.1037/0033-295X.115.2.502

[27] J. Klayman, J.B. Soll, C. Gonzalez-Vallejo, S. Barlas, Overconfidence: It Depends on How, What, and Whom You Ask, Organizational Behavior and Human Decision Processes. 79 (1999) 216-247. https://doi.org/10.1006/obhd.1999.2847

[28] J.B. Soll, J. Klayman, Overconfidence in Interval Estimates, Journal of Experimental Psychology: Learning, Memory, and Cognition. 30 (2004) 299–314. https://doi.org/10.1037/0278-7393.30.2.299

[29] A. Ancarani, C. Di Mauro, D. D'Urso, Measuring overconfidence in inventory management decisions, Journal of Purchasing and Supply Management. 22 (2016) 171-180. https://doi.org/10.1016/j .pursup.2016.05.001

[30] A. Grant, J. Franklin, P. Langford, The Self-Reflection and Insight Scale: A new Measure of Private Self-Consciousness, Social Behavior and Personality. 30 (2002) 821-836. https://doi.org/10.2224/sbp.2002.30.8.821

[31] L.F. Sasse-Werhahn, C. Bachmann, A. Habisch, Managing Tensions in Corporate Sustainability Through a Practical Wisdom Lens, Journal of Business Ethics. 163 (2020) 53-66. https://doi.org/10.1007/s10551-018-3994-z

[32] P. Goodwin, Improving the voluntary integration of statistical forecasts and judgment, International Journal of Forecasting. 16 (2000) 85-99. https://doi.org/10.1016/S0169-2070(99)00026-6

[33] J.M. Sargeant, K.V. Mann, C.P. van der Vleuten, J.F. Metsemakers, Reflection: a link between receiving and using assessment feedback, Advances in Health Sciences Education. 16 (2009) 399-410. https://doi.org/10.1007/s10459-008-9124-4